2013

Bioinformatics Approach to Probe Protein-Protein Interactions: Understanding the Role of Interfacial Solvent in the Binding Sites of Protein-Protein Complexes;Network Based Predictions and Analysis of Human Proteins that Play Critical Roles in HIV Pathogenesis.

Mesay Habtemariam
Virginia Commonwealth University

Bioinformatics Approach to Probe Protein-Protein Interactions: Understanding the Role of Interfacial Solvent in the Binding Sites of Protein-Protein Complexes; Network Based Predictions and Analysis of Human Proteins that Play Critical Roles in HIV Pathogenesis.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

By Mesay Habtemariam

B.Sc. Arbaminch University, Arbaminch, Ethiopia 2005

Advisors: **Glen Eugene Kellogg, Ph.D.**

Associate Professor, Department of Medicinal Chemistry & Institute For Structural Biology And Drug Discovery

Danail Bonchev, Ph.D., D.SC.

Professor, Department of Mathematics and Applied Mathematics, Director of Research in Bioinformatics, Networks and Pathways at the School of Life Sciences Center for the Study of Biological Complexity.

Virginia Commonwealth University

Richmond, Virginia

May 2013

ኃይልን በሚሰጠኝ በክርስቶስ ሁሉን እችላለሁ::

ፊልጵስዩስ 4፡13

I can do all this through God who gives me strength.

Philippians 4:13

Acknowledgment

A major research project like this is never the work of anyone alone. The contributions of many different people, in their different ways, have made this possible. I would like to extend my appreciation especially to the following.

My first and sincere appreciation goes to Dr. Glen E. Kellogg, Ph.D., my graduate advisor for all I have learned from him and for his continuous help and support in all stages of this thesis. Indeed, without his guidance, I would not be able to put the topic together. I would also like to thank him for being an open person to ideas, and for encouraging and helping me to shape my interest and ideas.

I offer my sincerest gratitude to my advisor, Dr. Danail G. Bonchev, Ph.D., who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way.

Dr. J.Neel Scarsdale, Ph.D. for serving as a member of my graduate student committee.

My gratitude is also extended to Dr. Allison A. Johnson, Ph.D., and Dr. Herschell S. Emery, M.Ed., Ph.D., for their assistance and guidance in getting my graduate career started on the right foot and providing me with the foundation for becoming a Bioinformatician.

I would like to gratefully and sincerely thank Mr. Mostafa Ahmed, Ph.D. student for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at Virginia Commonwealth University. His mentorship was paramount, I thank you.

I would like to thank Hardik Parikh, Ph.D candidate, who as a good friend, was always willing to help and give his best suggestions.

Many thanks to Dr. Philip Mosier, Chenxiao Da, Sreedevi Chandrasekaran, and the remain ISBDD members for their input, valuable discussions and accessibility.

I would also like to thank my family, especially my mother and father for always believing in me, for their continuous love and their supports in my decisions. Without whom I could not have made it here.

Finally, I Thank the Almighty God!!

Table of Contents:

LIST OF TABLES

LIST OF FIGURES

Abstract

BIOINFORMATICS APPROACH TO PROBE PROTEIN-PROTEIN INTERACTIONS: UNDERSTANDING THE ROLE OF INTERFACIAL SOLVENT IN THE BINDING SITES OF PROTEIN-PROTEIN COMPLEXES; NETWORK BASED PREDICTIONS AND ANALYSIS OF HUMAN PROTEINS THAT PLAY CRITICAL ROLES IN HIV PATHOGENESIS.

.By Mesay Habtemariam, M.Sc.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2013

Advisors: GLEN EUGENE KELLOGG, Ph.D.

Associate Professor, Department of Medicinal Chemistry & Institute For Structural Biology and Drug Discovery

DANAIL BONCHEV, Ph.D., D.SC.

Professor, Department of Mathematics and Applied Mathematics, Director of Research in Bioinformatics, Networks and Pathways at the School of Life Sciences Center for the Study of Biological Complexity.

The thesis work contains two projects under the same umbrella. The first project is to provide a detailed analysis on the behavior of interfacial water molecules at protein-protein complexes, in this case focusing on homodimeric complexes, and to investigate

their effect with respect to different residue types. For that reason the homodimeric data-set, which includes high-resolution (≤ 2.30 Å) X-ray crystal structures of 252 (140 Biological & 112 Non-biological) protein complexes was chosen to explore fundamental differences between interfaces that Nature has “engineered” vs. compared to interfaces found under man-made conditions. The data set was comprised of 5391 water molecules where a maximum of 4 Å from both interfacing proteins. Our analysis is applied a suite of modeling tools based on HINT, a program for hydropathic analysis developed in our laboratory. HINT is based on the experimental measurement of the hydrophobic effect. The second project is designed to explore various means of suppressing the expression of human genes that play critical role in HIV pathogenesis. To achieve this aim, a data set of Affymetrix Human HG Focus Target Array, which measures the expression levels of HIV seronegative and seropositive individuals in human PBMCs, was analyzed with Pathway Studio 9.0 software. This work gives insight into the elucidation of the important mechanisms of human proteins interactions in HIV seropositive individuals and their implications. Hence, we found the kind and types of microRNAs that are suppressing the human genes which have great role for HIV replication in a cell.

CHAPTER 1

GENERAL BACKGROUND

1.1 Protein-Protein Interactions

Animal and plant cells are filled with macromolecules that interact in a multitude of ways. Most of these interactions are transitory and trivial, but a few lead to the development of functionally relevant assemblies through the specific recognition of two partner molecules. Specific recognition is subject to strong positive biological selection, whereas the short-lived interactions undergo no selection or, more likely, they undergo a negative selection to prevent the formation of combinations that would harm the cell. Most of the macromolecules that exist in cells are proteins and their interactions with other proteins have many different chemical and physical bases. First, many biological processes are carried out, or regulated, through the interactions between preformed protein complexes [1]. The importance of such interactions in biology has made the protein recognition process an area of considerable interest. Second, many biological functions involving the formation of protein-protein complexes with finite lifetimes are formed between polypeptide chains of different sequences.

The two different types of complexes are homodimeric complexes, which are formed between two or more identical polypeptide chains and are usually symmetric, and hetrodimeric complexes, which are formed between different chains. In order to fully appreciate these biological associations, it is important to distinguish between the different types of complexes when analyzing the intermolecular interfaces that occur

within them [2]. The subunits that form homodimeric complexes are not found in nature as stable structures inside the cell, and the complex formation occurs concurrently during the folding process. On the other hand, the subunits of hetrodimeric complexes are regularly, but not constantly, independently stable inside the cell and they interact with each other to carry out a specific function in the cell.

It is worth noting that the stability of protein-protein complexes depends on the physiological conditions and the complex's surrounding environment [3]. For a large number of reasons it has been of a great interest over the past two decades to examine and understand the difference between various classes of protein-protein interfaces. Protein-protein interfaces have been subjected to many structural and computational analyses [4]. The interfaces of homodimeric complexes have greater numbers of interface residues and H-bonds than heterodimer interfaces, which means the density of hydrogen-bonds per residue is greater for heterodimer interfaces [2].

A large number of computational analyses of protein-protein interfaces have focused on what can be learned from the sequences and folding of the interacting proteins. Using only the amino acid composition of a protein-protein complex, Ofran and Rost were able to statistically predict interface classes correctly in up to 100% of the cases [5]. In fact, prediction tools using only protein expression information can often predict the complex type even in the absence of a 3D structure [6-8].

A more basic distinction can be drawn between protein-protein associations that occur in nature vs. those that are a consequence of experimental factors. One example of this, which is particularly important, is the difference between protein-protein

interactions that are biologically relevant and those that are a consequence of crystallization. In chapter two of this thesis, we are focusing on this difference.

It has been reported that support vector machine-based classification can be used to differentiate biological interactions from non-biological (crystal packing) contacts and differentiate obligate interactions from non-obligate [9]. In one algorithm, called NOXclass, the authors utilized six different attributes: interface area, ratio of interface area to protein surface area, amino acid composition of the interface, correlation between amino acid compositions of the interface and the overall protein-protein surface, interface shape complementarity, and conservation of the interface. NOXclass is reported to achieve 91.8% accurate classifications based on a leave-one-out cross-validation procedure [9]. Other support vector machine-based classifiers to predict protein-protein interface types have fared more poorly [10].

1.2 Role of water studies in protein-protein interface

The role of water molecules at protein-protein interfaces has been seeing increasing attention due to water's significant and varied contributions to protein-protein binding mechanisms [11]. The most basic role of a water molecule is to bridge polar interactions that are either too distant or energetically unfavorable. But, water molecules are also important even when they are displaced! They were found to be crucial in predicting hot spots (residues accounting for disproportionate binding free energy) in protein-protein complexes due to the water-entropy effect, which is a consequence of the hydrophobic effect [12]. Similarly, a recent review of polyproline recognition by protein-protein interaction domains showed that combining hydrophobic interactions

with strong networks of water-mediated hydrogen bonds is a mechanism that has been exploited repeatedly to favor the adaptability and plasticity of different families of proteins [13].

Despite the extensive studies on understanding the differences among protein-protein interface types, the role of water in classifying these interfaces has received less attention. In most cases, water is an important protein structural feature that may add plenty of information to the protein interfacial definition [14]. Sonavane and Chakrabarti examined the cavities between subunits in homodimeric and heterodimeric complexes, respectively, and their hydration states, and found that the fraction of water molecules possessing a direct hydrogen bond with both subunits was 37% and 51%, in homodimeric and heterodimeric complexes, respectively, and that the fraction with hydrogen bonds to neither subunit was 10% and 5%, respectively[15]. However, this analysis was not performed on protonated and H-bond optimized structures; thus no information on the quality of the reported hydrogen bonds could be provided. Nevertheless, this study also quantified the role of water molecules in neutralizing the destabilizing effect of like-charges on the two interacting subunits [15].

In a protein-water-protein interface model of a nested-ring, an atom re-organization method was used to detect hydration trends and patterns between biological and non-biological interfaces [16]. According to this model, biological interfaces are found to be drier than the non-biological interfaces. That research organized atoms at the same burial level in each tripartite protein-water-protein interface into a ring. Then, the rings of an interface are ordered with the core atoms placed at the center of the structure to form a nested-ring topology. Based on this topology, Li et al.

[16] found that water molecules on the rings of an interface are generally configured in a dry-core-wet-rim pattern with a progressive level-wise solvation towards to the rim of the interface and that his solvation trend becomes sharper when counter ions are separated [16]. Their analysis was based solely on solvent-accessible surface area (SASA) of water molecules and their contact distances and used B-factors for further investigation. It should be noted, however, that we previously did not find B-factors to be useful for the prediction of water conservation [17].

In a previous study from our group [18], a data-set of 179 high resolution (< 2.30 Å) X-ray crystal structures that was composed of mainly biological hetero-protein-protein complexes with all hydrogens in optimized orientations, we reported that of the 4741 interfacial water molecules: a) 21% were involved in (bridging) interactions favorable with both proteins; b) 53% were favorably interacting with only one protein; and c) 26% had no favorable interactions with either protein. This trend was shown to independent of the crystallographic resolution, which supports the assertion that the majority of even the water molecules unfavorable with respect to both proteins are not crystallographic assignment errors or artifacts. It was also shown that the interactions of water molecules with residue backbones are consistent for all classes, accounting for 21.5% of all interactions, and that interactions with polar residues are significantly more common for bridging waters, while interactions with non-polar residues dominate the last group. Water molecules that interact favorably with both proteins stabilize on average the protein-protein interaction by (-0.46 kcal mol^{-1}), but overall, the average contribution of a single water molecule to the protein-protein interaction energy is unfavorable (+0.03 kcal mol^{-1}). Interestingly, analysis of the waters without favorable

interactions with either protein suggests that this is a conserved phenomenon: 42% of these waters have SASA ≤ 10 Å and are thus largely buried within the protein-protein interface, and 69% of these are within predominantly hydrophobic environments. Such water molecules may have an important biological purpose in mediating protein-protein interactions [19].

Based on the fundamental and intriguing results of our previous study [18], we have expanded our work to investigate a larger and more themed dataset in order to better understand roles of water molecules in forming the interfaces of biological and non-biological complexes. This thesis describes a detailed analysis of interfacial water molecules found in 252 X-ray crystal structures of protein-protein complexes extracted from the RCSB Protein Data Bank [20]. Of the 252 X-ray crystal structures, 140 are from biological homo-protein protein complexes while the other 112 structures are from non-biological protein-protein interfaces, (i.e., protein-protein interactions that are believed to be formed only under crystallographic conditions).

In these studies, hydrophobicity is the major factor that stabilizes protein-protein association; thus, their complementarity plays a selective role in defining which proteins may associate [21]. Bahadur, et al. [4] suggests that understanding and being able to predict non-biological associations could be key to discriminating inappropriate protein-protein binding that leads to disease [1].

In chapter two of this thesis, we seek to understand protein-protein interactions by answering a few questions about their interfaces: Are the waters at non-biological interfaces playing the same role as those at biological interfaces? Are they energetically

favorable or unfavorable for each residue type? Which association type, biological or non-biological, relies more on water mediated interactions with backbone atoms? Is there a significant difference on the total energetic contribution of water for biological interfaces and non-biological interfaces?

1.3 Protein-Protein Interaction Networks

Biological interactions of proteins with other proteins are variable in their nature and are heterogeneous, both spatially and temporally [22]. The varied natures of protein-protein interactions (PPIs) make the construction and analysis of biological network models a thought-provoking topic in the field of biological complexity as we attempt to represent their underlying systems. Network views of PPIs are undoubtedly powerful when a detailed view of a given subsystem is analyzed [23]. To accurately signify the interactions in the proteome, Hakes et al. [22] suggested two points: first, network properties need to be understood, and second, reasonably complete datasets are required. In this way we can compose detailed information concerning the nature of interactions, including the specific functional implications of each interaction to ensure the connection between network analysis and biological understanding.

The availability of large-scale protein-protein interaction data has led to the recent popularity in the study of protein interaction networks. Just as an immense amount of available sequence data has made it possible to attain an overview of the genome, it is hoped that this newly available interaction data will allow an analogous view of the interactome. The prospect of proposing biological conclusions from this

network structure is part of what makes protein-protein interaction data so fascinating and significant.

Exploring protein-protein interactions on a more macroscopic level leads to the use of network analyses. In chapter three of this thesis, we seek to understand protein-protein interactions networks by answering a few questions about their networks: Do MicroRNAs play great role as post-transcriptional regulators to influence human proteins that play critical role in HIV pathogenesis? Do the designated human proteins have significant interactions with other human proteins? How do we connect the implications of each human protein that play critical role in HIV pathogenesis with other human proteins and MicroRNAs, from their networks?

Chapter 2

WATER MOLECULES AT PROTEIN-PROTEIN INTERFACES

This chapter describes a general analysis of homodimeric protein-protein interfaces in 252 high resolution (better than 2.3 Å) X-ray crystal structures of protein-protein complexes extracted from the RCSB protein Data Bank[20] and an in-depth assessment of the role of water molecules at biological and non-biological protein-protein interfaces.

2.1 Material and Methods

2.1.1 Data Set

2.1.2 Data set I:

The protein-protein complex data set was obtained from an informational portal to a biological macromolecular structure database called the RSCB Protein Data Bank [20] by applying search filters for several structural criteria. The selection of these complexes were based on : 1) if the structure contains two chains of the same protein that have a minimum chain length of at least 100 amino acids, and 2) a structure with X-ray resolution 2.3 Å or better. Finally, 252 structures (Appendix I) were randomly selected from this set for analysis.

2.1.3 Data set II:

These datasets derived from the previous total dataset (Appendix I), are categorized to Biological (Appendix II) and non-biological (Appendix III) protein complexes based on ‘REMARK 350’ of each PDB file. The biological homo-dimer protein complexes are complexes that occur naturally, presumably for a biological purpose, whereas the non-biological interfaces are protein-protein interactions that are believed to be formed only under crystallographic conditions.

2.2 Methods

After the list of proteins were organized, hydrogen atoms were added to each protein and these were minimized (Tripos forcefield, with Gasteiger-Hückel charges and distance-dependent dielectric) to a gradient of 0.01 kcal mol^{-1} $Å^{-1}$, while the non-hydrogen atoms were treated as an aggregate using Sybyl 8.1[24]. Then, the HINT modeling system was used for all further structural analysis.

2.2.1 Hydropathic INTeractions (HINT):

HINT is a novel empirical molecular modeling system for de novo drug design and protein or nucleic acid structural analysis [25]. The scoring used by the HINT model for biomolecular interaction is based on experimental logP for 1-octanol/water partitioning. HINT simultaneously accounts for enthalpy, entropic and solvation contributions to biological association [16] [25-27]. HINT also calculates 3D hydrophatic interaction maps that are uniquely instructive for understanding biomacromolecular structure: substrate /inhibitor/drug binding to proteins and nucleotides, protein subunit interactions and protein folding [26].

Several studies have been done via HINT and most of them resulted in an output that showed, in an intuitive way, the types and quality of the binding interactions between the ligand and the receptor [17]. In general, the software is useful: 1) to estimate LogP for modeled molecules or data files, 2) numerically and graphically evaluate binding of drugs or inhibitors into protein structures and scores docked ligand orientations, 3) to construct hydrophatic (lock and key) complementarity maps that can be used to predict an ideal substrate from a known receptor or protein structure, and 4)

to evaluate/predict effects of site-directed mutagenesis on protein structure and stability [28].

For our analysis, we have given close attention for the HINT score, Rank, H-Bond Score, Acid/Base Score, Hydrophobic Score, Acid/Acid Score and Base/Base Score results of HINT. The combination of HINT score and Rank gives the 'Water Relevance" of each interaction, which was designed as a global metric for describing the conservation of water between unliganded and ligand-bound states in protein complexes [29], but which is here extended to protein-protein complexes.

2.2.2 Hydropathic Analysis

In our study of protein-protein complexes in this study, each model contains two proteins and an array of solvent molecules. Each was analyzed with HINT [25] by computing intermolecular scores between the proteins and the interfacial solvent arrays. The HINT score (H_{TOTAL}) is a double sum over all atom-atom pairs of the product (b_{ij}) of the hydrophobic atom constants (a_i, partial log $P_{octanol/water}$) and atom solvent accessible surface areas (S_i) for all interacting atoms, mediated by a function of the distance between the atoms:

$$H_{TOTAL} = \sum_i \sum_j b_{ij} = \sum_i \sum_j (a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij}) \quad (1)$$

where R_{ij} is a simple exponential function, e^{-r} [25], r_{ij} is an adaptation of the Lennard-Jones function [30-31], and T_{ij} is a logic function assuming +1 or -1 values, depending on the polar (Lewis acid or base) nature of interacting polar atoms. HINT parameters and controls were as in previous studies [17][32-33]: partition calculations were performed with the "dictionary" method for the proteins with 'essential hydrogens',

where polar hydrogens are treated explicitly and non-polar hydrogens are 'united' with their parent non-polar heavy atom; the HINT option that corrects the S_i terms for backbone amide nitrogens by adding 30 $Å^2$ was used in this study to improve the relative energetics of inter- and intramolecular hydrogen bonds involving these nitrogens. Water molecules are partitioned as a "solvent set" with analogous HINT parameters. Previous work [34-35] has suggested that approximately 500 HINT score units correspond to -1.0 kcal mol^{-1} of free energy.

Each crystallographically observed orientation of water was optimized by an exhaustive protocol [28] that maximizes the HINT score with respect to its surrounding environment by evaluating its interactions with a "receptor" created from atoms within 6.0 Å of it. For water molecules, this optimization rewards hydrogen bond and acid/base interactions while penalizing acid/acid and base/base interactions and those with hydrophobic entities on either of the two protein surfaces. Hydropathic interaction analysis was then performed with HINT for each of the optimized water molecules with respect to the two proteins with which it interacts. The resulting data were tabulated by frequency and strength of interactions with each amino acid residue type. In cases where a water molecule had significant interactions (>|10| HINT score units, approximately |0.02| kcal mol^{-1}) with more than one residue on a protein, that water's count was fractionally distributed to interacting residues based on the absolute values of the relative HINT scores for those residues that interact with it, i.e.,

$$W_i = \sum_n \{ |A_i^c| / \sum_i |A_i| \} \quad (2)$$

where A_i^c are the interaction HINT scores by residue type (i) interacting with water n. Similarly, the fractions of interactions with interfacial water molecules arising from backbone and sidechain atoms were calculated by weighted counts with A_i^c representing the interaction HINT scores by i, separated into c = sidechain or c = backbone subsets. Heat maps for frequency and interaction scores and map clustering were calculated and drawn with R [36].

2.2.3 Rank Algorithm

Rank represents the weighted number of potential hydrogen bonds for each water molecule with respect to a pseudo-receptor of atoms from the target molecule(s) surrounding the water. Rank is calculated as:

$$\text{Rank} = \sum_n \{ (2.80\ \text{Å}/r_n) + [\sum_m \cos (\theta_{Td} - \theta_{nm})]/6 \} \quad (3)$$

where r_n is the distance between the water's oxygen and the target's heavy atom n (n is the number of interaction hydrogen bond donor/acceptor (doneptor) targets up to a maximum of 4). This is scaled relative to 2.8 Å, the presumed ideal hydrogen bond length. θ_{Td} is the optimum tetrahedral angle (109.5°) and θ_{nm} is the angle between targets n and m (m = n to number of valid targets). The algorithm thus allows a maximum number of 4 doneptor targets (≤ 2 donors and ≤ 2 acceptors). To properly weight the geometrical quality of hydrogen bonds, targets that have an angle less than 60° with respect to other (higher quality) targets are rejected [28].

2.2.4 Relevance

Relevance is a synthesis of HINT score and Rank [29]. Specifically,

$$\text{Relevance} = \{ P_R(|W_R| + 1)^2 + P_H(|W_H| + 1)^2 \} / \{ (|W_R| + 1)^2 + (|W_H| + 1)^2 \} \quad (4)$$

where P_R is the percent probability for water conservation based on Rank and P_H the probability based on HINT score. W_R and W_H are the weights for these probabilities, respectively. The values for P_R, P_H, W_R and W_H are as shown in Figure 2 of [29]. This relationship was derived with the expectation that water molecules with Relevance ≥ 0.5 would be conserved and those with Relevance < 0.5 would be non-conserved because the waters observed in unliganded proteins and analyzed in developing the training set were, by their nature, binary – either conserved and present in the ligand-bound complex or non-conserved and absent in the complex.

2.3 Results and Discussion

One of the main and unique abilities of water is to provide two hydrogen-bond acceptor sites and two donor sites. Thus, it can effectively bridge in every way possible [8]. In general, there are three distinct roles for waters at protein-protein interfaces: bridging i.e., having significant interactions with both proteins; non-bridging, i.e., having significant interactions with only one of the two proteins; or simply trapped without significant interactions with either protein.

The result and discussion in this chapter is divided into two specific objectives:

1. A thorough analysis of water molecules at homo-dimer protein-protein interfaces and a comparison with previous results [18] observing the water molecule contribution in a homodimeric data-set.

2. A detailed report on the role of water molecules between biological and non-biological protein complex interfaces and also quantifying the interfacial water molecules for each residue type.

2.3.1 Homodimeric Analysis

2.3.1.1 The water Relevance metric

We applied the Relevance algorithm to the set of water molecules at homodimeric protein-protein interfaces to understand to recognize their roles in these complexes. Interface water molecules are those that are a maximum of 4 Å from atoms in both proteins. The homodimeric dataset includes 252 proteins, comprised of 5391 unique water molecules. Each complex has a number of interfacial water molecules, between 1 and 469 waters or an average of 75 at the protein-protein interface. Figure 2.1 illustrates the percentage of water molecules for different Relevance classes. These classes correspond to how many proteins the water is Relevant with respect to. For all water molecules in this study 19% of them have Relevance class two, whereas 29% and 52% are in the zero and one Relevance classes respectively.

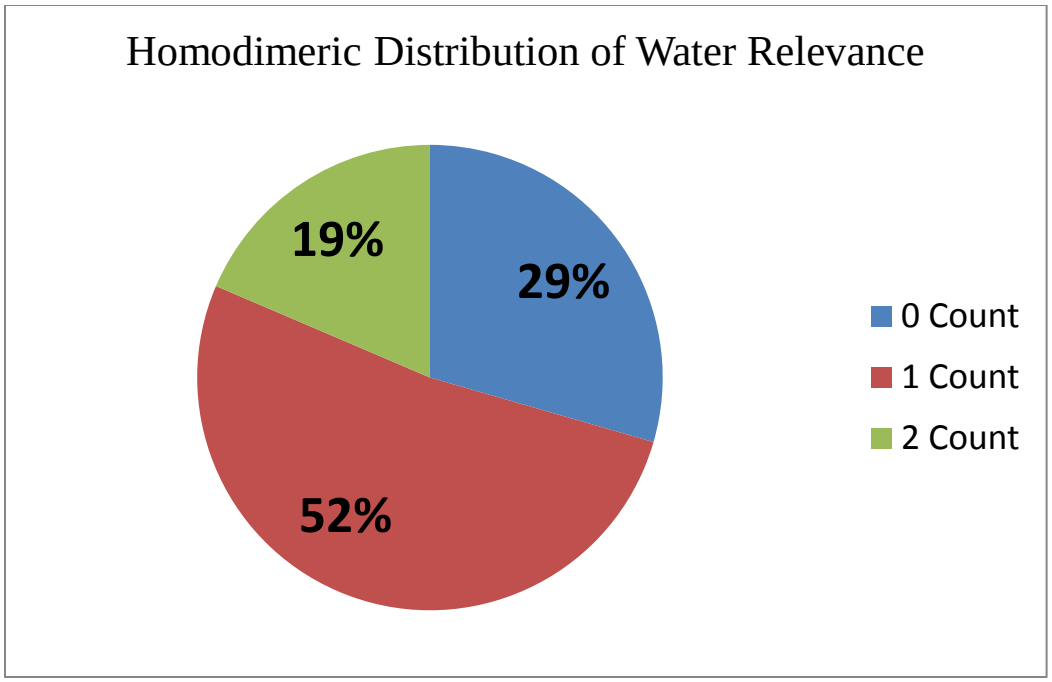


Figure 2.1 Distribution of interfacial water molecules by Relevance in the homodimeric protein-protein interfaces.

This result suggests that just one-fifth of the waters that are found at the protein-protein interface are truly bridging by binding with both proteins, while more than half of the waters are strongly associated with only one of the two proteins. On the other hand, nearly a one-third are not Relevant with respect to either protein. The water molecules that bind to only one protein can provide steric constraints for the protein association but they do not provide significant favorable energetic contribution to the association [18].

2.3.1.2 Residue Preference for Interface H2O

As defined above, water Relevance [29] is a metric from the combination of Rank [28] and HINT score [25]. To further understand the role of water molecules at protein-protein interfaces, we applied the Relevance algorithm [29] to categorize the interfacial waters by amino acid residue types.

The frequency and HINT scores of water molecules are tabulated by interaction counts (Table 2.1). These weighted counts are calculated as $\Sigma n\{ |A_i| / \Sigma_i |A_i| \}$, where A_i are the interaction HINT scores by residue type (i) interacting with water n and the HINT scores are averaged two ways: first, over all waters in the set or Relevance subset, and second, by frequency (weighted count) of that residue type in the set or Relevance subset. As it was shown in earlier [18], the more polar residues, in particular Aspartate (Asp = 11.6%) and Glutamate (Glu =10.8 %), appear most often in interactions involving water at protein-protein interfaces. Cystine (Cys), even though it is a polar amino acid, is most rarely (i.e., 0.4%) found. However, the non-polar aliphatic hydrophobic residues: Glycine(Gly), Isoleucine(Ile), Valine(Val), Proline(Pro), Alanine(Ala), and Leucine(Leu) showed a prevalently negative HINT score, but frequencies of 4.6%, 5.4%, 6.4%, 6.9%, 8.0% and 8.9%, respectively.

These results are in qualitative agreement with our earlier study [18] for all water interactions between residues at protein-protein interfaces [18]. In fact, the percentage variations are very much similar for the zero, one and two Relevance classes. For instance, waters having Relevant interactions with both proteins, the polar acidic Asp and Glu amino acids, as well as the polar basic Lysine(Lys),Histidine(His) and Arginine (Arg) amino acids, exhibit frequency ranges from 4.1 % to 19.8%.

Table 2.1. Frequencies and HINT scores of water molecules at homodimeric protein-protein interfaces with respect to interacting amino acid residues.

	All Waters			Water Relevant to 0			Water Relevant to 1			Water Relevant to 2		
Residue Type	wtd. Count	Average Hint Score		wtd. Count	Average Hint Score		wtd. Count	Average Hint Score		wtd. Count	Average Hint Score	
		For All	For Type		For All	For Type		For All	For Type		For All	For Type
ALA	433	-38.81	-483.66	213	-68.24	-510.26	194	-32.80	-474.40	26	-8.78	-336.19
ARG	342	14.66	231.10	46	3.14	108.08	190	15.60	223.37	105	31.50	299.01
ASN	196	6.92	190.61	32	2.84	139.21	112	7.47	186.55	51	11.87	231.98
ASP	627	59.26	509.53	49	11.30	364.60	373	68.31	513.64	205	109.87	536.91
CYS	19	0.01	2.97	5	-0.37	-179.94	11	0.09	50.42	3	0.40	108.14
GLN	195	5.21	144.44	37	1.38	58.85	112	5.51	137.51	45	10.47	232.60
GLU	581	46.26	429.54	65	7.18	174.76	350	57.53	460.31	165	76.76	465.05
GLY	247	-9.45	-206.50	76	-12.96	-269.72	132	-9.81	-208.37	39	-2.87	-74.71
HIS	131	4.80	197.97	20	1.96	158.06	111	5.39	135.96	42	7.66	181.66
ILE	294	-28.64	-526.04	96	-30.58	-507.09	159	-30.78	-540.62	38	-19.59	-512.75
LEU	480	-43.18	-484.64	189	-56.10	-472.87	227	-39.32	-485.21	65	-33.44	-516.89
LYS	262	-1.96	-40.45	62	-1.55	-39.66	142	-1.50	-29.58	58	-3.93	-67.85
MET	162	-12.91	-430.55	65	-16.61	-527.29	79	-12.16	-430.04	18	-9.11	-512.90
PHE	95	2.13	120.07	21	1.15	87.08	51	2.29	124.71	23	3.22	139.86
PRO	373	-29.11	-421.18	195	-51.73	-422.31	154	-23.70	-430.00	24	-8.35	-354.21
SER	188	-4.45	-127.40	52	-5.04	-153.15	98	-5.09	-145.14	38	-1.72	-377.59
THR	274	-17.68	-348.04	98	-21.68	-351.27	137	-17.02	-347.52	39	-13.19	-341.69
TRP	44	1.16	143.44	13	1.58	194.14	22	1.01	127.14	9	0.94	109.60
TYR	105	4.17	213.84	27	4.22	245.31	56	4.42	221.38	22	3.36	154.87
VAL	343	-31.90	-500.87	154	-51.70	-531.78	167	-28.91	-486.13	22	-8.86	-397.02

2.3.1.3 Distribution of HINT scores for water molecules

The average HINT score for the waters in the entire data set is -12 ($\Delta G \sim +0.024$ kcal mol^{-1}); in other way, the total average HINT score (i.e., the sum) for all residue types and for all waters is -73.51($\Delta G \sim +0.15$ kcal mol-1). Table 2.1 lists the HINT score values for each of the twenty amino acid types, by averaging over all waters in the data set and by averaging over all waters interacting (by weighted count) with that residue type.

The distribution of HINT score for all water molecules ranges significantly (Figure 2.2). Very informative distributions, however are observed between water molecules Relevant to zero, one and two proteins. In Figure 2.3, for Relevance zero, the HINT scores are mainly found in a range less than zero, and these waters have average HINT scores of -348.9 (+0.69 kcal mol^{-1}). In the case of Relevance one (Figure 2.4), the average HINT scores is -36.4 (+0.07 kcal mol^{-1}). Finally, the Relevance to two HINT scores predominantly have positive values of up to 500 (Figure 2.5) with an average HINT score of 202 (-0.40 kcal mol^{-1}).

Negative HINT scores are unfavorable while positive scores are favorable. Thus, the waters Relevant to neither protein are dominated by interactions with non-polar hydrophobic residues (i.e., Ala, Ile, Leu, Pro, Thr and Val) while for the waters Relevant to both proteins, the polar residues (Arg, Asp, Lys, His and Glu) dominate the interactions.

Figure 2.2 Histograms illustrating distribution of HINT scores for all water molecules in data set

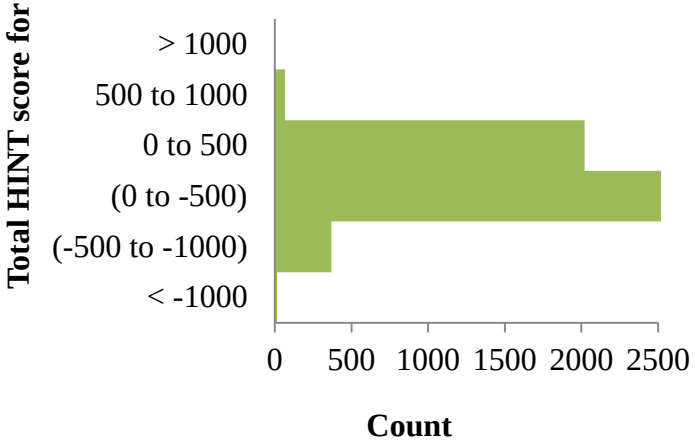


Figure 2.3 Histograms illustrating distribution of HINT scores for water molecules with Relevance to neither protein

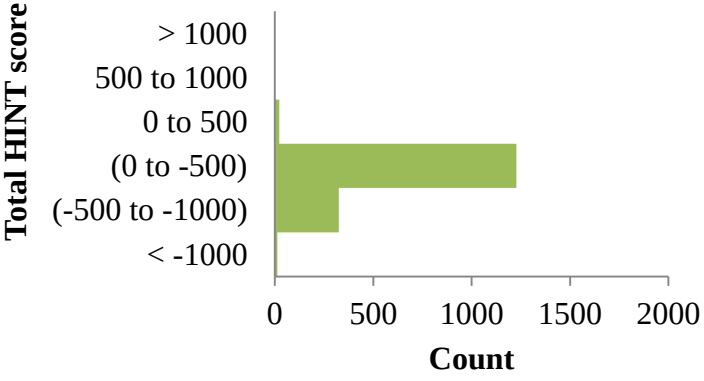


Figure 2.4 Histograms illustrating distribution of HINT scores for water molecules with Relevance to one protein

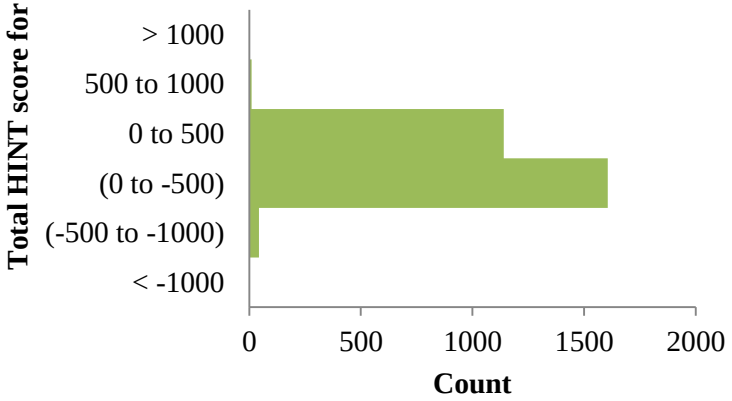
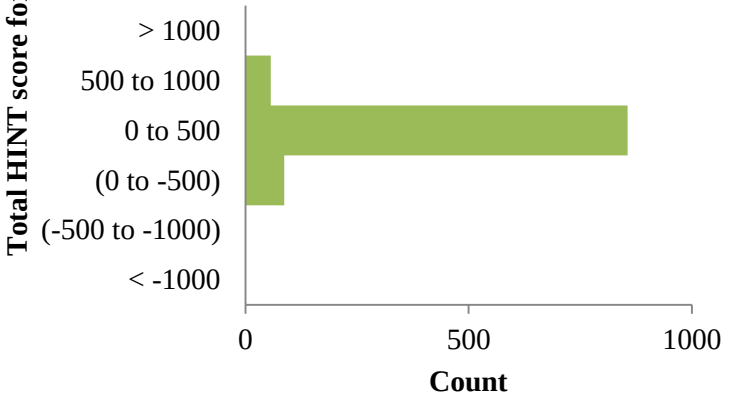


Figure 2.5 Histograms illustrating distribution of HINT scores for water molecules with Relevance to both protein



2.3.2 Biological vs Non-biological Analysis

2.3.2.1 The role of water molecules on Biological vs. non-biological interfaces

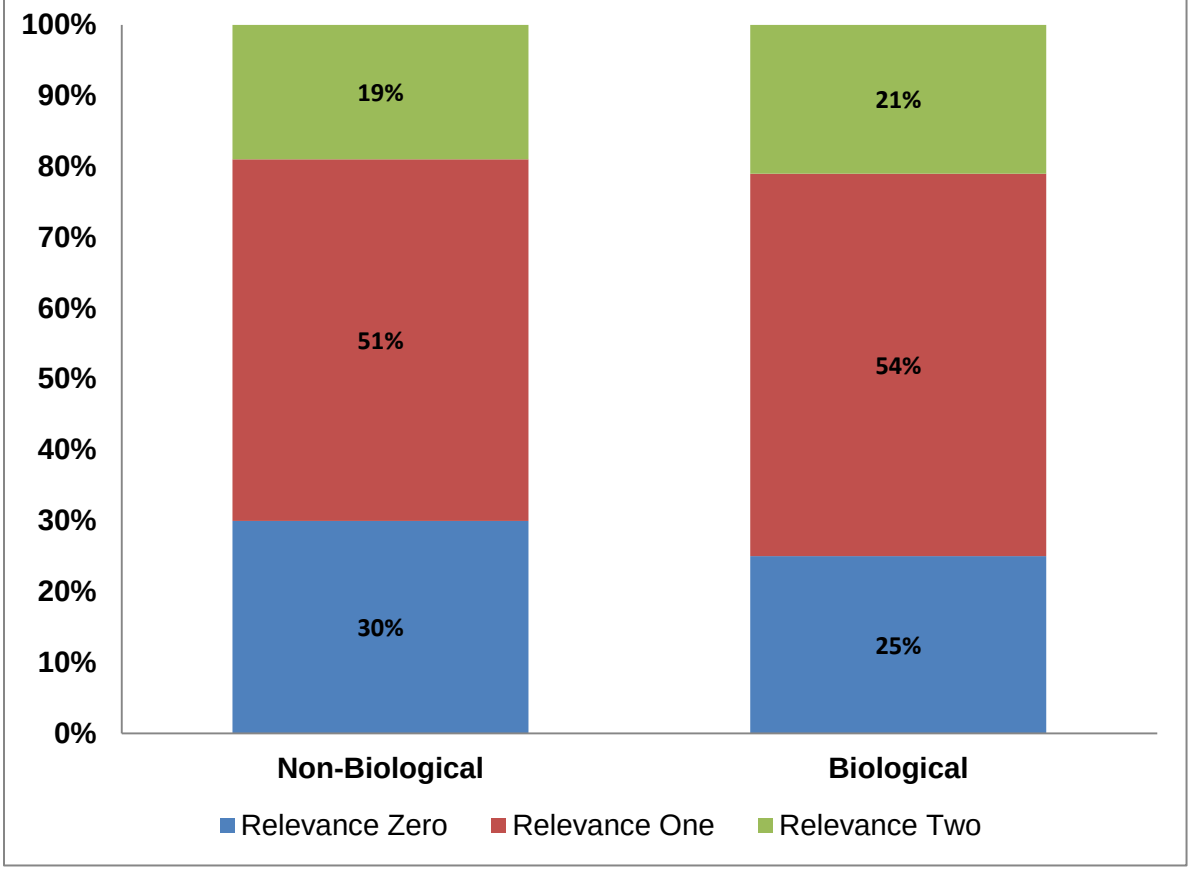
The contribution and effect of water molecules in life are countless. They are the key molecules of life: from complex interdependent ecosystems to being a key component of nearly every biological reaction and interaction on the molecular scale. While protein-protein interactions are a topic of increasing relevance in the quest for new approaches to treat disease, much of the mechanism of this machinery of life is poorly understood, not the least of which are the roles of the many discrete water molecules observed at structurally characterized protein-protein interfaces. As it is known that a significant fraction of protein-protein interactions observed in X-ray crystal structures are not biologically relevant, but are, in fact, a consequence of the crystallographic lattice, the first question of some significance is: what are the characteristics of such “biological” and “non-biological” interfaces and are there differences in the roles that water molecules play in these two cases?

In previous studies of waters in the interface between interacting proteins, researchers have generally relied on interatomic distances in non-protonated crystallographic models to mark interactions between waters and proteins. This approach, however, often poorly represents the complex and subtle energetics and geometric preferences of hydrogen bonding [28].

Based on the classification scheme described above, for waters Relevant to zero, one or two proteins, we examined the biological and non-biological datasets. First,

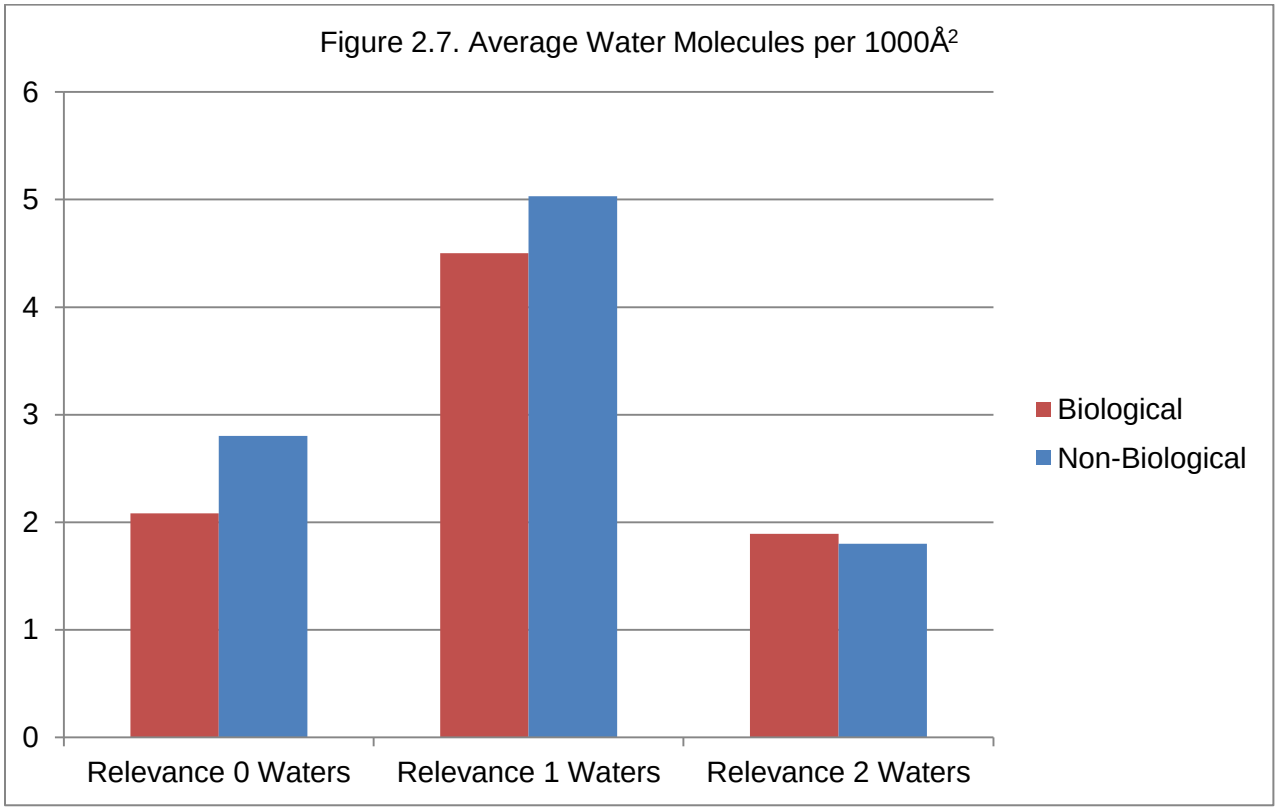
the results for the homo-oligomeric biological set of complexes are virtually identical to our previous result for hetero-oligomeric biological complexes [18]. As shown in Figure 2.6, although the differences are not dramatic and not statistically significant, with both types of interfaces having more or less the same percentages, there is a tendency for the biological interfaces to have a larger fraction of “stabilized” waters and the crystallographic interfaces to have a larger fraction of “non-stabilized” waters.

Figure 2.6. Distribution of Water Relevance in Biological & Non-Biological data set



.

Second, the average interface area of Biological interface was found to be significantly higher ($p < 0.05$) than that of the Non-Biological interfaces in our data set (2660 $Å^2$ and 1510 $Å^2$, respectively), which is in agreement with previous reports [9][14]. However, when examining the "wetness" of the two interfaces (average number of water molecules per 1000 Å), it was found that there are 8.5 and 9.6 water molecules per 1000 Å on average for Biological and Non-Biological interfaces, respectively. Normalizing the number of waters in each relevance class by the interface areas of the complexes yields the result plotted in Figure 2.7. The difference in water density between biological and non-biological interfaces was found to be statistically significant for the case of Relevance zero waters ($p < 0.05$).



2.3.2.2 Residue Preference for waters

Next, the HINT scores and water preferences for each residue type were calculated for both datasets (see Table 2.2). Residues have similar preferences when interacting with water in the two types of interfaces with the exceptions of His, Phe and Tyr, where the preferences for interaction with water are as much as doubled in the biological dataset. These residues, along with Cys, Met and Trp, are rare on protein surfaces [14], and their increased presence in water-mediated biological interactions would seem to be purpose-driven, while only random for the non-biological interface cases. Also fairly dramatic is the increase in preference for Lys interacting with water at the non-biological interfaces (6.69% vs. 4.91%), which may be due to Lys's high charge density and sidechain flexibility, which allows it to opportunistically interact with water molecules over a fairly large area.

To explore these observations more quantitatively, in terms of energetics, HINT score values were calculated for each water molecule in the dataset. The average water at a biological interface has a HINT score of 20 (-0.04 kcal mol^{-1}), while at the non-biological interface it is -14 (+0.03 kcal mol^{-1}). Thus, as before [18], the *average* water is essentially meaningless. Furthermore, for each of the twenty amino acid types, these scores were summed were averaged in two ways, first by averaging over all waters in the data set, and second by averaging over all waters interacting (by weighted count) with that residue type. The first average, over all waters, reveals the magnitude and the nature of interaction; i.e., whether it is energetically favorable (HINT score > 0) or unfavorable (HINT score < 0) for each residue with water. The latter average, weighted instead by the frequency of that particular water-residue interaction,

represents the score that would be expected if a “typical” water molecule interacted with only that residue, and thus reveals the specific benefits or costs of interacting with each residue type. Tracking with the frequency results noted above, the average HINT scores for His and the hydrophobic residues Phe and Tyr are, for the biological dataset, more than twice those of the non-biological dataset. These are three of the four residue types capable of π-stacking (the other being Trp), which is one of the reasons that they are less commonly found on protein surfaces. The weighted HINT scores (Table 2.2) reveal the unsurprising result that hydrophobic residues have unfavorable interactions with water, while polar residues, with hydrogen bonding functional groups on their side chains, have favorable interactions with water. The trends in differences between water molecules found at biological interfaces compared to those found at non-biological interfaces mirror, for the most part, what was seen in the unweighted average scores.

Table 2.2. The frequency of water-residue interactions and an average HINT score for all waters and residue type

All Waters						
Residue Type	Residue Preference of water (%)		Average HINT score for all waters		Average HINT score for residue type	
	Biological	Non-Biological	Biological	Non-Biological	Biological	Non-Biological
ALA	6.19	5.69	-28.08	-23.47	-453.41	-412.51
ARG	5.89	5.54	16.66	15.01	282.73	271.02
ASN	5.36	6.09	7.84	8.35	146.15	137.04
ASP	13.32	12.2	63.67	53.29	477.96	436.76
CYS	0.51	0.77	0.27	0.99	53.99	129.07
GLN	4.75	4.79	6.32	5.51	133.12	115.02
GLU	13.07	14.77	50.7	53.22	387.79	360.12
GLY	5.87	5.66	-13.18	-10.32	-224.4	-182.13
HIS	2.6	1.29	4.84	2.58	185.83	200.89
ILE	3.61	4.02	-19.29	-21.94	-534.52	-545.89
LEU	5.8	5.46	-28.46	-25.93	-490.47	-474.65
LYS	4.91	6.69	2.05	1.41	41.8	21.03
MET	2.12	1.53	-9.86	-6.49	-463.84	-423.47
PHE	1.59	0.9	1.64	0.37	102.81	41.35
PRO	5.14	5.98	-22.8	-27.24	-443.13	-455.43
SER	5.39	6.16	-3.69	-3.9	-68.45	-63.3
THR	5.87	5.65	-15.75	-16.49	-268.49	-291.83
TRP	0.82	0.78	1.11	1.28	135.69	163.91
TYR	2.4	1.53	4.6	1.19	191.26	77.9
VAL	4.75	4.48	-25.45	-21.46	-535.94	-478.84

We performed similar analyses on subsets of water molecules based on their Relevance classes. Our intent was to discern differences between the interfaces in biological and non-biological complexes based on the roles of the waters trapped in those interfaces. Tables 2.3, 2.4 and 2.5 show the residue preferences and the average HINT scores for Relevance Zero, Relevance One and Relevance Two waters, respectively.

In the case of Relevance Zero waters, generally, the more hydrophobic and π-stacking residues (with the exception of Ile) are found preferentially in water interactions at biological interfaces, where Nature may have engineered a role for these non-stabilized water molecules. Likewise, the most polar residues, Arg, Asp, Glu and Lys, along with Ser, are present in more interactions with water at the non-biological interfaces. These residues are often found at surfaces and their involvement in water-bridged non-biological interactions with other proteins may be simply opportunistic.

Relevance zero water has the same average HINT score (-285, +0.55 kcal mol^{-1}), dominated by hydrophobic-polar interactions (see Figure 2.10 & Figure 2.11) for both biological and non-biological interfaces. One observation of note in Table 2.3 is the differences between residue-weighted HINT scores for biological and non-biological waters for interactions with Trp and Tyr. Trp interacts favorably with water (-0.31 kcal mol^{-1}) at biological interfaces, but is energetically neutral with respect to water at non-biological interfaces. Tyr interacts favorably with water (-0.25 kcal mol^{-1}) at biological interfaces, but unfavorably (+0.20 kcal mol^{-1}) at non-biological interfaces.

Figure 2.8. Average HINT interaction score for Main Chain atoms

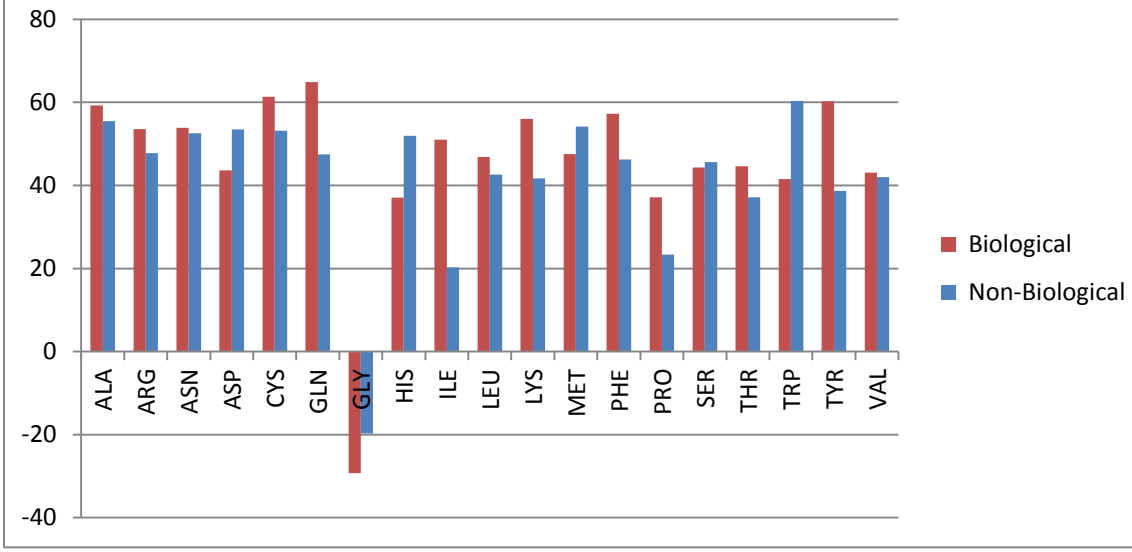


Figure 2.9. Average HINT interaction score for Side Chain atoms

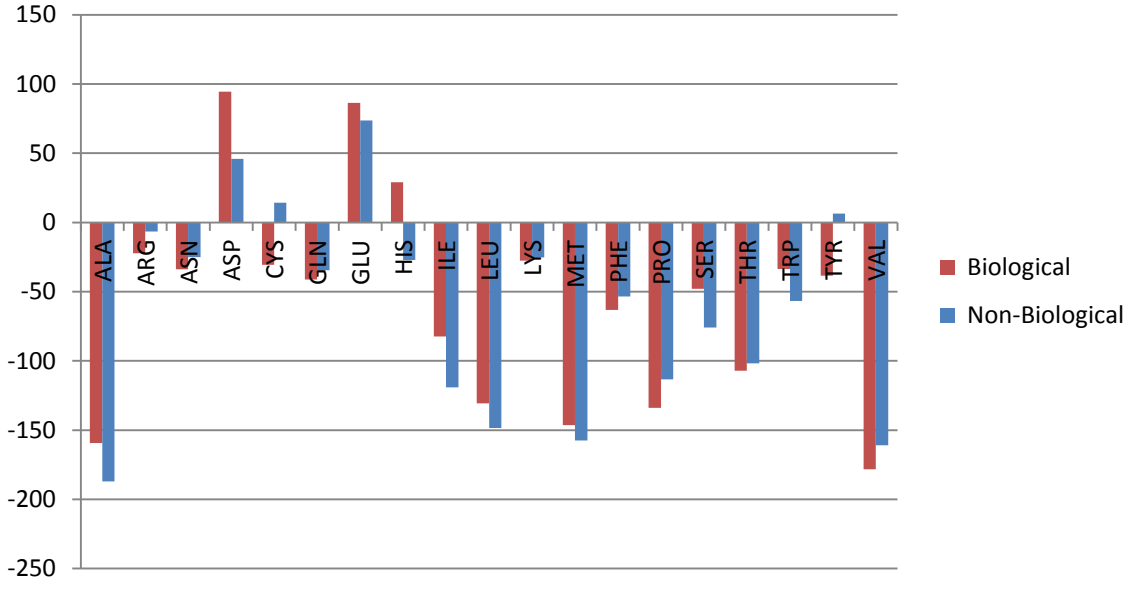


Table 2.3. The frequency of water-residue interactions and an average HINT score for waters Relevance to zero

Relevance Zero						
Residue Type	Residue preference of water (%)		Average HINT score for all waters		Average HINT score for residue type	
	Biological	Non-Biological	Biological	Non-Biological	Biological	Non-Biological
ALA	10.94	9.85	-50.88	-42.68	-464.68	-433.29
ARG	2.9	3.43	3.16	4.49	108.76	130.92
ASN	4.24	3.87	3.27	1.58	77.08	40.73
ASP	4.05	4.77	11.4	10.35	281.06	216.85
CYS	0.55	0.54	-0.17	-0.16	-30.11	-29.51
GLN	3.85	3.94	2.52	1.53	65.51	38.71
GLU	6.34	7.54	6.55	1.58	103.29	20.95
GLY	6.88	6.73	-18.18	-14.79	-264.16	-219.81
HIS	1.78	0.89	2.36	1.98	132.33	221.64
ILE	4.83	6.44	-23.47	-35.77	-485.96	-555.18
LEU	9.08	8.82	-43.52	-41.66	-479.22	-471.82
LYS	4.15	5.92	-4.21	-7.21	-101.46	-121.69
MET	3.67	2.23	-19.04	-11.78	-518.39	-529.15
PHE	1.73	0.62	0.71	0.56	40.97	89.45
PRO	10.52	11.08	-44.03	-50.48	-418.27	-455.57
SER	4.54	7.26	-5.29	-8.4	-116.47	-115.74
THR	8.72	7.7	-25.59	-25.36	-293.34	-329.02
TRP	0.64	0.53	1.01	0	158.12	0.88
TYR	2.07	1.2	2.68	-1.25	129.3	-104.24
VAL	8.47	6.6	-45.98	-29.85	-542.71	-452.27

There are some, mostly insignificant, differences between the water roles at biological vs. non-biological interfaces in the Relevance One cases (Table 2.4). The surprisingly limited energetic contribution of Lys, for both types of interface, is an artifact of its dual nature – i.e., having a very polar amine functional group and a long hydrophobic chain; thus, on average, its two contributions cancel out.

The average HINT scores for Relevance One waters for the biological and non-biological datasets are -33 (+0.06 kcal mol^{-1}) and -5 (+0.01 kcal mol^{-1}), respectively. This is seemingly a little unexpected; however, analysis of the types of interactions (Figure 2.12) reveals that Relevance One waters in the biological dataset have a larger contribution of unfavorable hydrophobic-polar interactions.

Table 2.4. The frequency of water-residue interactions and an average HINT score for waters Relevance to one.

Relevance One						
Residue Type	Percentage preference of water		Average HINT score for all waters		Average HINT score for residue type	
	Biological	Non-Biological	Biological	Non-Biological	Biological	Non-Biological
ALA	5.73	4.53	-27.09	-18.79	-472.77	-414.89
ARG	5.88	6.42	15.89	17.78	270.38	276.69
ASN	5.6	6.53	9.02	9.68	161.05	148.17
ASP	14.56	14.34	68.61	62.86	471.38	438.24
CYS	0.53	0.93	0.44	1.55	83.01	166.79
GLN	4.99	4.64	6.92	4.54	138.69	97.71
GLU	14.06	16.39	57.4	63.51	408.33	387.43
GLY	5.93	5.68	-12.49	-9.85	-210.5	-173.35
HIS	2.42	1.31	4.62	2.46	191.22	187.77
ILE	3.79	3.67	-22.19	-19.74	-585.87	-538.12
LEU	5.7	4.48	-29.56	-21.86	-518.36	-488.41

LYS	4.58	6.39	-0.22	0.39	-4.87	6.07
MET	1.75	1.56	-8	-5.38	-457.94	-344.05
PHE	1.43	1.05	1.8	0.24	125.96	22.47
PRO	4.18	4.51	-20.04	-20.4	-479.84	-452.72
SER	5.67	5.56	-4.16	-3.84	-73.32	-69.09
THR	5.74	5.29	-16.18	-15.98	-282.16	-301.97
TRP	0.76	0.83	0.67	1.4	87.86	168.42
TYR	2.2	1.61	4.57	2.68	208.08	167.22
VAL	4.51	4.27	-24.02	-22.01	-532.05	-514.9

The differences between the two types of interfaces for Relevance Two waters are also not dramatic (Table 2.5). Relevance Two waters in biological interfaces have higher preferences (~ 2-fold) for His, Met, Phe and Tyr, which again may be related to the relative unlikelihood of these residues being found on regions of a protein surface not engineered for biologically-relevant interaction.

Table 2.5. The frequency of water-residue interactions and an average HINT score for waters Relevance to two.

Relevance Two						
Residue Type	**Percentage preference of water**		**Average HINT score for all waters**		**Average HINT score for residue type**	
	Biological	**Non-Biological**	**Biological**	**Non-Biological**	**Biological**	**Non-Biological**
ALA	1.9	2.28	-4.32	-5.88	-227.75	-258.02
ARG	9.38	6.45	34.21	24.06	364.9	373.1
ASN	6.04	8.41	10.07	15.43	166.6	183.49
ASP	20.85	18.09	111.33	94.99	534.03	524.99
CYS	0.41	0.67	0.37	1.27	88.83	190.08
GLN	5.16	6.53	9.16	14.43	177.59	221.1
GLU	18.32	21.79	84.48	106.67	461.01	489.47
GLY	4.56	3.94	-9.19	-4.53	-201.58	-114.99

HIS	4.03	1.84	8.25	3.87	204.89	210.37
ILE	1.74	1.15	-6.98	-6.11	-402.39	-531.19
LEU	2.27	2.84	-8.22	-12.19	-362.2	-429.56
LYS	6.66	8.74	15.15	17.77	227.65	203.29
MET	1.31	0.35	-4.02	-1.15	-307.19	-326.55
PHE	1.86	0.94	2.29	0.46	123.41	48.52
PRO	1.42	1.95	-5.36	-9.18	-378.13	-471.1
SER	5.67	6.1	-0.65	3.02	-11.53	49.45
THR	2.9	3.38	-3.27	-3.89	-112.73	-115.11
TRP	1.18	1.04	2.37	2.95	201.15	284.62
TYR	3.32	1.82	6.87	0.98	207.25	53.55
VAL	1.04	1.7	-5.39	-6.72	-515.72	-395.28

.

Biological and non-biological interfaces have nearly the same average interaction type scores for waters with Relevance to zero, one and two proteins (Table 2.6 or Figure 2.10) and (Table 2.7 or Figure 2.11).

Table 2.6 (Biological) Average interaction type scores for waters with Relevance to zero,one and two proteins

Water Relevance	H-Bond and Acid/Base	Acid/Acid and Base/Base	Hydrophobic/Polar
0	571	-390	-454
1	1014	-684	-349
2	1317	-872	-230

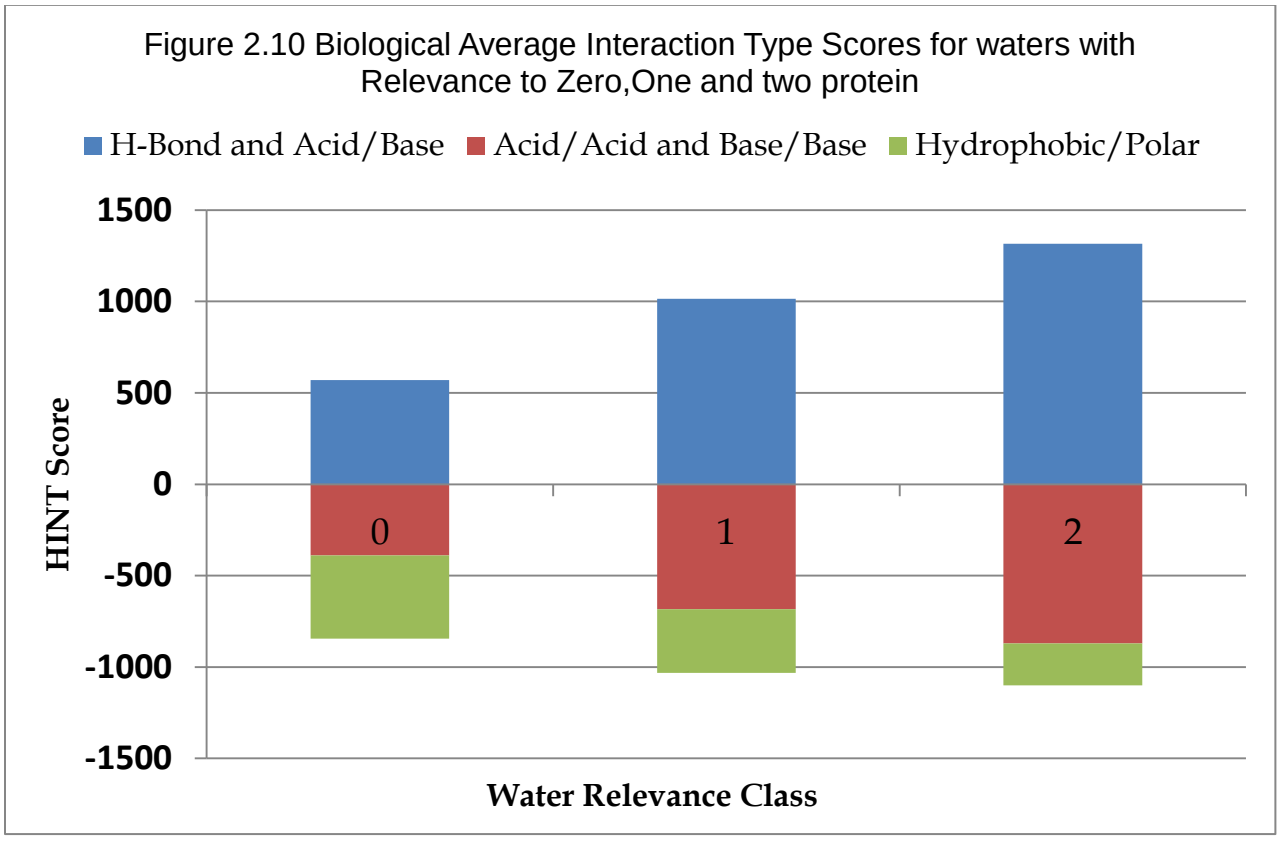


Figure 2.10 Biological Average Interaction Type Scores for waters with Relevance to Zero,One and two protein

Table 2.7 (Non-biological) Average interaction type scores for waters with Relevance to zero, one and two

Water Relevance	H-Bond and Acid/Base	Acid/Acid and Base/Base	Hydrophobic/Polar
0	489	-338	-426
1	980	-664	-308
2	1343	-897	-233

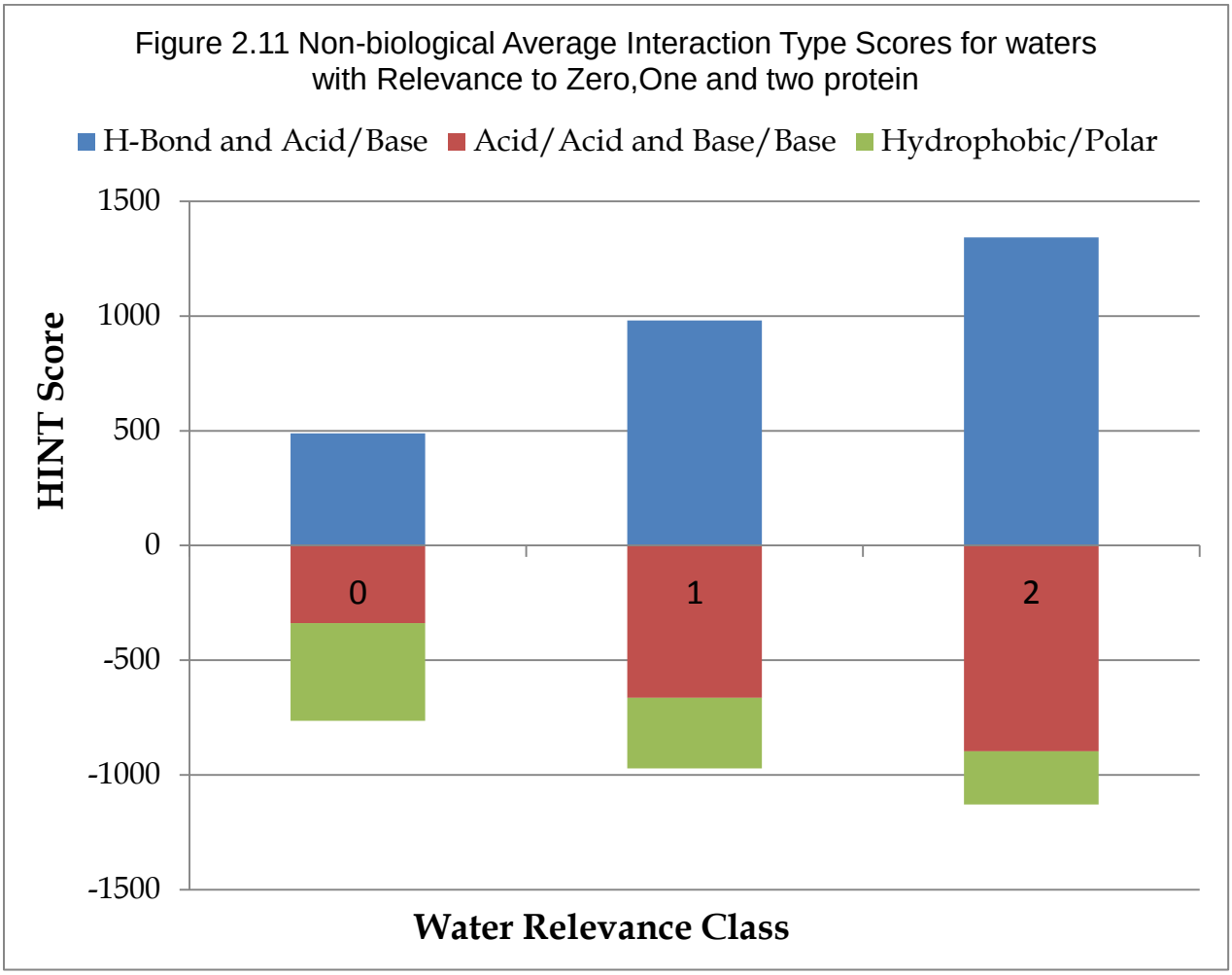


Figure 2.11 Non-biological Average Interaction Type Scores for waters with Relevance to Zero,One and two protein

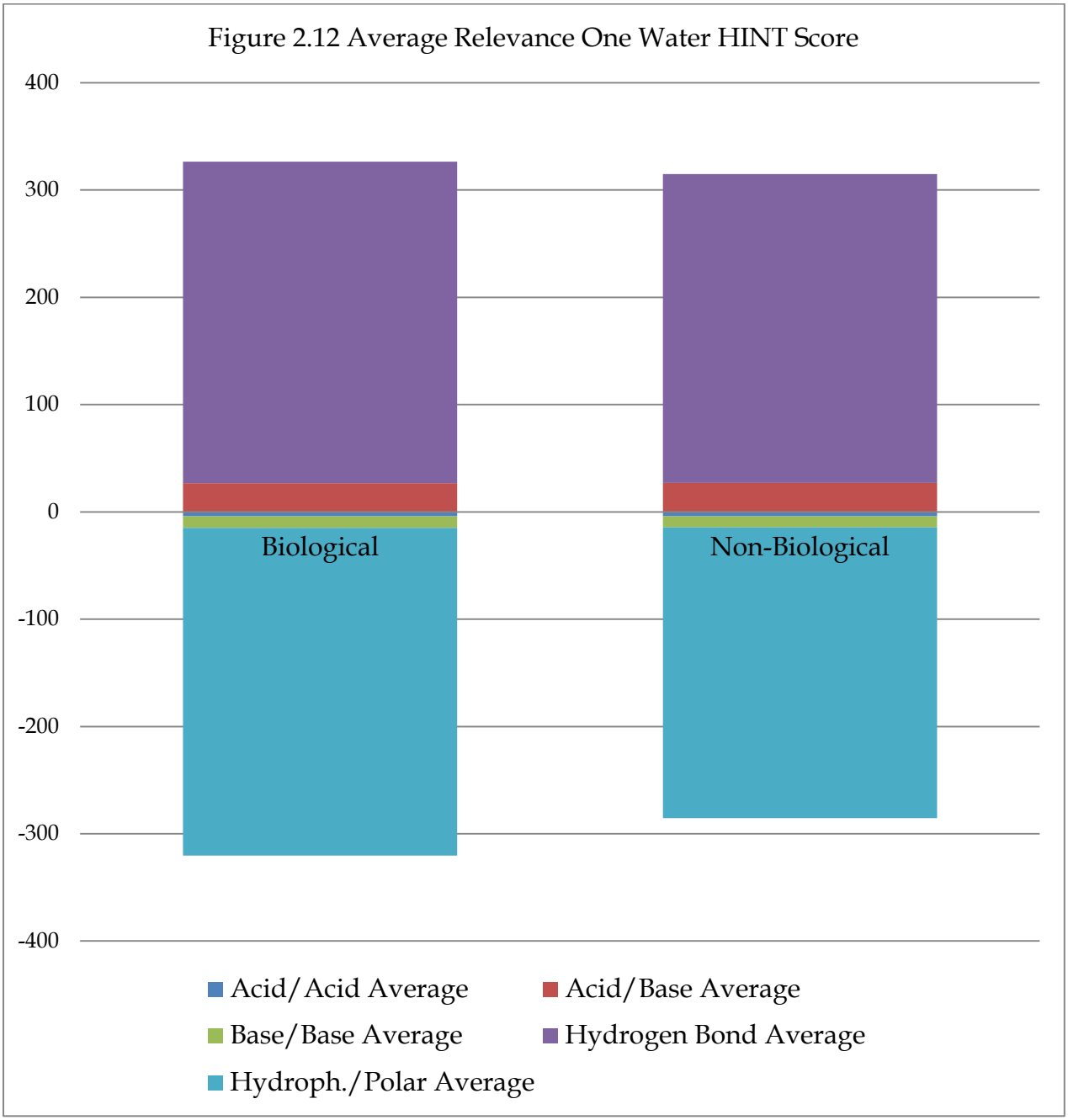


Figure 2.12 Average Relevance One Water HINT Score
400
300
200
100
0
-100
-200
-300
-400
Biological
Non-Biological
Acid/Acid Average
Acid/Base Average
Base/Base Average
Hydrogen Bond Average
Hydroph./Polar Average

2.3.2.3 Backbone and Sidechain Preferences for Interfacial water

We next separated the residue's contributions to interactions with water into those arising from main chain (backbone) atoms, i.e., C, O, N and CA, and their bonded hydrogens, and those arising from the sidechain atoms. This may help us understand better the differences between residue interactions in the two datasets. Our previous analysis [18] revealed that the average interaction score for a water molecule with a backbone atom is favorable, while the average interaction with sidechain atoms is unfavorable, although the identity of the sidechain plays an obvious major role. We expected to see differences in modes of interaction with water for residues in biological vs. non-biological complexes. Figure 2.8 shows the HINT score averaged by number of water molecules for the backbone atoms in the biological and non-biological datasets. It appears that with the exception of Asp, Gly, His, Met, Ser and Trp, all the other residues have higher average HINT scores in biological complexes than those of the non-biological complexes. Of those, Lys, Pro, Gln, Gly, Ile and Tyr are significantly higher ($p<0.05$), which explains the discrepancies found in Table 2.2. Figure 2.9 shows the HINT score averaged by number of water molecules for the Biological and Non-biological datasets for the side chain atoms only. Surprisingly, polar residues such as Asp, Glu and His have better water mediated HINT scores in biological complexes. In particular, His has an unfavorable average score in non-biological interfaces in contrast to biological interfaces where His has a favorable water mediated average HINT score. Also, hydrophobic residues like Ala, Ile, Leu, Met and Trp have a better average HINT scores in biological complexes; however, only Asp and His were found to be significantly higher ($p<0.05$).

2.3.2.4 Residue Pair Preferences for All Water

Figures 2.13 and 2.14 show the heat maps for HINT scores for all residue pairs interacting with waters, normalized by weighted frequency. These illustrate the propensity and energetics of water molecules "bridging" between the specified residue types; the deeper blue cells represent more favorable situations, e.g., Glu-H_2O-Glu or Asp-H_2O-Tyr, while the deeper red cells represent highly unfavorable situations, e.g., Ile-H_2O-Val. The heat maps appear to be similar with only subtle differences between biological and non-biological cases. Clustering of these maps (Figures 2.15 and 2.16), however, more clearly highlights differences: in the biological interfaces (Figure 2.15), all charged polar residues (with the exception of His, which has a pK_a of around 6.0) are dramatically separated from the other residues. Cys, Met, Phe and Trp, are in a second distinct cluster whose commonality is difficult to understand, while the remaining twelve residue types are in very flatly defined clusters with mixed electronic properties for the member residues. However, in non-biological interfaces (Figure 2.16) the clean distinction between charged and uncharged residues is no longer seen, as Asp, Glu, Arg and Lys cluster with the uncharged polar residues Asn and Ser, and surprisingly, Gly. The backbone of Gly contains both a good hydrogen bond acceptor (O) and donor (N+H) and is more exposed than the backbones of all other residues, but is also the most hydrophobic backbone – possessing a methylene at CA. The remaining thirteen residue types are distributed in two clusters that are even flatter than those observed for the biological interface case, but with similar content.

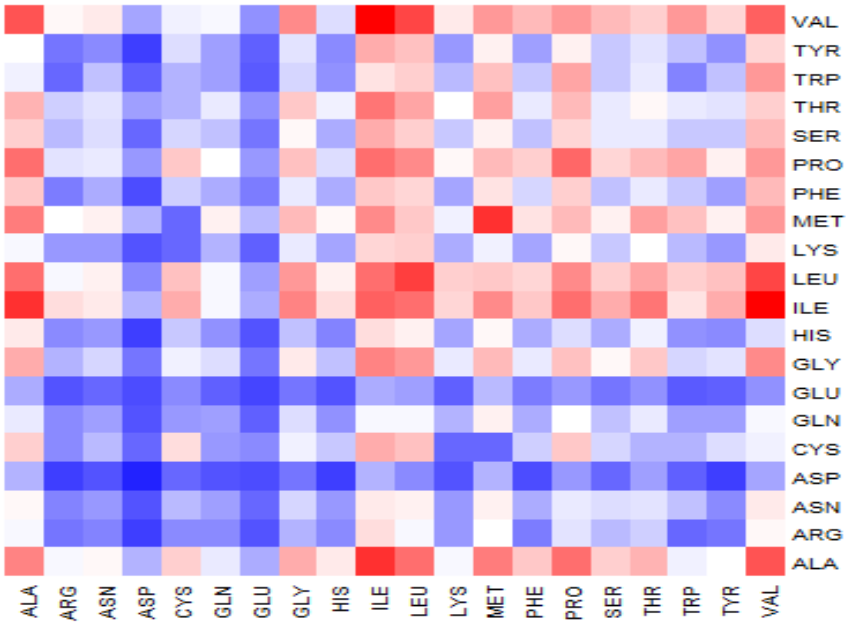


Figure 2.13 Heat maps depicting Res1-H2O-Res2 interactions for all water molecules found at Biological protein-protein interfaces.

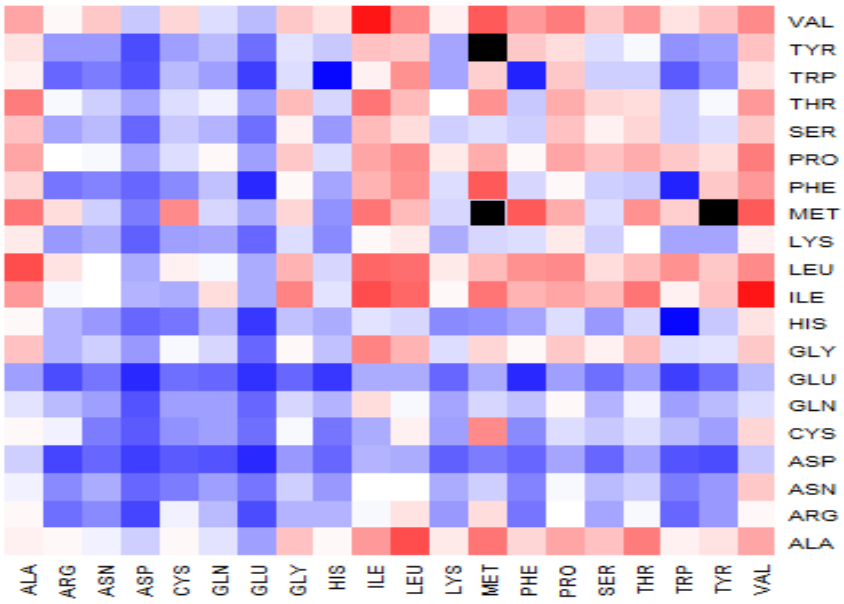


Figure 2.14 Heat maps depicting Res1-H2O-Res2 interactions for all water molecules found at non-biological protein-protein interfaces.

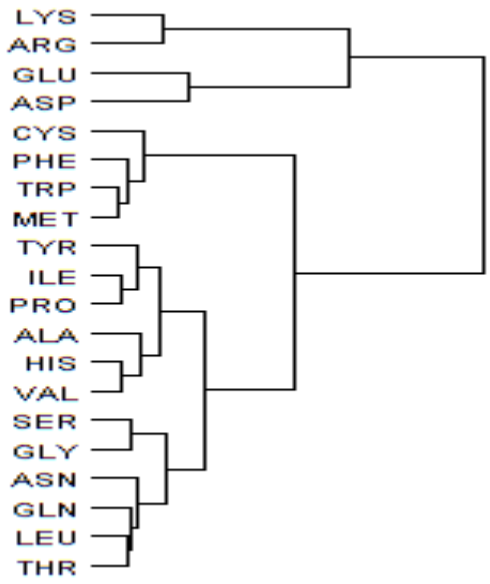


Figure 2.15 Dendogram indicating clustering of residues with respect to average HINT score (normalized by weighted count) in Biological Res1-H2O-Res2 interactions for all waters

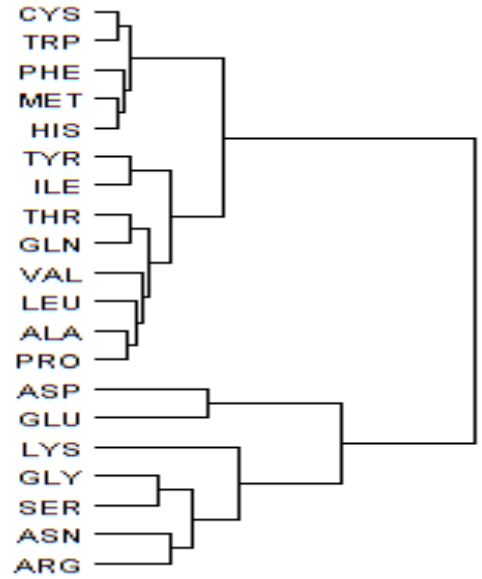


Figure 2.16 Dendogram indicating clustering of residues with respect to average HINT score (normalized by weighted count) in non-biological Res1-H2O-Res2 interactions for all waters

2.3.2.5 Crystallization and Water

The conditions of crystallization force protein chains to bind together in conformations/arrangements that may not always be favorable. The conditions of crystallization affect both biological and non-biological complexes and results in bringing the binding partners closer together. Such associations may be unfavorable for some of the water molecules involved (i.e., Relevance Zero waters); however, they lead to a globally favorable energetic minimum for the whole complex. In this case, the total energetic contribution of water ranges between -2.35 and 3.67 kcal mol^{-1} (average -2.10 kcal mol^{-1}) for biological interfaces and between –1.38 and 3.07 kcal mol^{-1} (average -1.46 kcal mol^{-1}) for non-biological interfaces (see Table 2.10).

Table 2.8. Average Total Energy of Waters for Protein-Protein Interfaces by Relevance

Average Total energy of waters for Protein-protein Interfaces by Relevance		
	Average Energy (Kcal mol-1)	
	Biological	Non-biological
All waters	2.1	1.46
Relevance Zero	3.67	3.07
Relevance One	0.87	0.27
Relevance Two	-2.35	-1.38

In an analysis based solely on SASA, contact distance and B-factors of water molecules, Li et al. [16] used a tripartite protein-water-protein interface model and a nested-ring atom re-organization method to detect hydration trends and patterns between obligate, non-obligate and non-biological interfaces [16]. According to their model, biological interfaces are found to be drier than the non-biological interfaces. Our

analysis in the present work indicates that, although biological interfaces are more hydrophobic than non-biological interfaces, more importantly they are better designed to accommodate small polar molecules like water by engaging more of their hydrophobic residues' backbone atoms in less unfavorable interactions with water molecules. Some of the polar residues interact better with water molecules in biological interfaces than in non-biological interfaces. Non-biological interfaces, on the other hand, are not designed to interact together, which often leads to unfavorable interaction with water – even though they are more polar than non-biological interfaces. For a specific residue type, Tyr is more frequently found involved with water at biological interfaces compared to non-biological interfaces, but more specifically, Tyr interacts favorably with water in the Relevance Zero biological interfaces cases but unfavorably in the non-biological interface cases. Nature's designated role for Tyr in protein-protein associations is clearly subverted in non-biologically relevant associations.

2.4 Conclusions

This analysis of water molecules at biological and non-biological protein-protein interfaces has revealed new information about the structure of these interfaces. Our analysis was anchored by the HINT free energy forcefield and the Relevance metric. The former characterizes the types and qualities of interactions between the interface waters and proteins, while the latter is a simple parameter that was previously shown to identify water molecules conserved/non-conserved in ligand binding sites [29]. This work on homodimeric complexes, differentiating between biological (largely obligate) and non-biological interfaces, is an extension of a previous study [18] of heterodimer complexes that were generally transiently formed.

First, from the perspective of water, there are surprisingly few differences between heterodimer and homodimeric datasets. The broadest classification scheme we employed, by Relevance, showed the same distribution of waters (within 1%) amongst the Relevance Zero, One and Two classes. This was somewhat surprising, as we expected that homodimeric formation would be using bridging waters more profitably than heterodimers. Clearly, water is more than ubiquitous in protein-protein systems: it is pervasive. Second, even the differences between biological interfaces and crystallographic (non-biological) interfaces are relatively modest at the Relevance class level: there are 5% more Relevance Zero waters at the non-biological interfaces, resulting in 3% and 2% fewer Relevance One and Zero waters, respectively, at non-biological interfaces. Again, this is somewhat surprising, as we expected a significantly larger fraction of Relevance Zero waters for the artificial constructs of crystallographic contacts, and a much larger fraction of Relevance Two waters for the obligate/biological interfaces where folding and association are more or less simultaneous, i.e., engineered. Third, looking much deeper into the differences by analyzing the roles of different residues at these interfaces revealed a few notable observations: i) non-biological interfaces are more polar than biological interfaces, yet there is better organized hydrogen bonding at the latter; ii) biological associations rely more on water-mediated interactions with backbone atoms compared to non-biological associations – an indication of engineering by Nature; iii) aromatic/planar residues play a larger role in biological associations with respect to water because these residues would not normally be found on the surface unless there was a planned role for them; and iv) Lys has a peculiar role: it is often found on protein surfaces with its main role apparently solvating

the protein, but because of its flexibility and reach, plays an out-sized role in forming non-biological interfaces as it can often find a direct or water-mediated hydrogen-bonding partner.

Chapter 3

MOLECULAR INTERACTIONS NETWORKS OF HUMAN PROTEINS THAT PLAY CRITICAL ROLES IN HIV PATHOGENESIS

3.1 Introduction:

One of the major goals in proteomics is to develop a complete description of the protein interaction networks that underlie cell physiology. Conventionally, a protein-protein interaction map is epitomized as a static network, where each node represents a protein and each edge represents a protein-protein interaction. These maps are called PPINs [37]. In reality, a PPIN is a dynamic entity because the functional state of the network depends on the expression of protein nodes [38]. In this chapter, we will first give a brief background about Human Immunodeficiency Virus (HIV), before we start describing the building of the interaction network of the human proteins that play critical role in HIV pathogenesis (HPPCR-HIV pathogenesis). HIV belongs to a class of viruses known as retroviruses. Retroviruses are viruses that contain RNA (ribonucleic acid) as their genetic material.

HIV has a small genome and therefore relies heavily on the host cellular machinery to replicate. Identifying which host proteins and complexes come into physical contact with the viral protein is crucial for a comprehensive understanding of how HIV rewires the host's cellular machinery during the course of infection [39-40]. After infecting a cell, HIV uses an enzyme called reverse transcriptase to convert its

RNA into DNA (deoxyribonucleic acid) and then proceeds to replicate itself using the cell's machinery [41].

Each day, HIV destroys billions of $CD4^+$ T cells in a person infected with HIV, eventually overwhelming the immune system's capacity to regenerate or fight other infections. When HIV infects a cell, the virus can hide within the cytoplasm (the jelly-like fluid that fills the cell) or integrate into the cell's genetic material (chromosome).Shielded from the immune system, HIV can lie dormant in an infected cell for months or even years. These cells serve as a latent reservoir of the virus [39-40].

A map of the physical interactions between proteins within a particular system is necessary for studying the molecular mechanisms that underlie the system. The analysis of interacting human and viral proteins has been successfully done using a variety of methods [40]; however, viral proteins can mimic native interfaces and thus interfere with binding events in host protein networks [42]. Also the knowledge of the set of interacting human proteins that play great roles in infectious diseases would greatly contribute to our understanding of the mechanisms of infection, and subsequently to the design of new therapeutic approaches [43]. We need, therefore, to look at a more complete presentation of protein interactions using common regulator and shortest path protein-protein networks.

In this report we aim to identify the association of the HPPCR-HIV pathogenesis proteins with other human proteins and microRNAs, systematically and quantitatively, using Pathway Studio Software, version 9.0. MicroRNAs play a major role as post-transcriptional regulators to influence a large proportion of genes in higher eukaryotes

[38]. First, we identified from literature resources 19 human proteins that play critical roles in HIV pathogenesis, including their accessory factors. By providing these testable protein data, we are able to define a reliable network of human protein-protein interactions. Then, we combine our selected proteins with a list of proteins generated from a gene expression data obtained using microarray experiments. Finally, we show evidence that most of the human proteins that have important roles in HIV pathogenesis are regulated by microRNAs in the PPINs.

3.2 Data and Methods

The first 19-Human proteins that play critical role in HIV pathogenesis (HPPCR-HIV pathogenesis) data-set is derived from previously published HIV-human PPIs and host factors implicated in HIV function [39]. Jager et al. [40] explored in detail the biological significance of viral and human proteins interactions to advance the structural modeling of viral and human PPIs.

We next analyzed a gene expression profile of HIV-1 patients and control samples to explicate the functional genomic relationships with other human proteins identified in the first data-set. The microarray data is derived from “Gene Expression Omnibus” (GEO) (www.ncbi.nlm.nih.gov/geo/) database that contains functional genomic data in Array- and sequence-based format.

Table 3.1. List of 19-HPPCR-HIV pathogenesis and their number of neighbors used to build the network.

Name	**Info**	**Description**
OTUD4	20 neighbors	OTU domain containing 4
CCR5	657 neighbors	chemokine (C-C motif) receptor 5
CCL5	1164 neighbors	chemokine (C-C motif) ligand 5
CXCR4	1150 neighbors	chemokine (C-X-C motif) receptor 4
CD4	2823 neighbors	CD4 molecule
IL10	2490 neighbors	interleukin 10
KAT5	290 neighbors	K(lysine) acetyltransferase 5
PPIA	298 neighbors	peptidylprolyl isomerase A (cyclophilin A)
CCR3	246 neighbors	chemokine (C-C motif) receptor 3
CCNT1	142 neighbors	cyclin T1
HIVEP1	16 neighbors	human immunodeficiency virus type I enhancer binding protein 1
HTATIP2	62 neighbors	HIV-1 Tat interactive protein 2, 30kDa
HTATSF1	25 neighbors	HIV-1 Tat specific factor 1
HIVEP2	53 neighbors	human immunodeficiency virus type I enhancer binding protein 2
VPRBP	31 neighbors	Vpr (HIV-1) binding protein
ITIH4	29 neighbors	inter-alpha (globulin) inhibitor H4 (plasma Kallikrein-sensitive glycoprotein)
TARBP2	34 neighbors	TAR (HIV-1) RNA binding protein 2
TARBP1	11 neighbors	TAR (HIV-1) RNA binding protein 1
AGFG2	8 neighbors	ArfGAP with FG repeats 2

3.2.1 Microarray Data

Protein-based microarrays have been shown recently to be promising tools for analyzing small amounts of samples, while yielding the maximum data on the cell's environment [44]. Microarray technology has become one of the indispensable tools that many biologists use to monitor genome-wide expression levels of genes in a given organism. A standard way of getting the data is by comparing expression of a set of genes from a cell maintained under particular conditions to the same set of genes from a reference cell maintained under normal conditions.

In this analysis we used the Affymetrix Human HG Focus Target Array that measures the expression levels of HIV seronegative and seropositive individuals in human PBMCs in vivo [45]. Ockenhouse et al. [45] took a total of 87 primary clinical samples consisting of human peripheral blood mononuclear cells (PBMC), including 12 seronegative samples from healthy control subjects, 22 seropositive samples from drug-naïve persons, 21 seropositive samples from persons who had received at least 1 antiretroviral drug regimen, and 32 seropositive samples from persons whose $CD4^+$ T cell counts either decreased or increased during the study period. Seropositive persons with differential changes in $CD4^+$ T cell counts may have received nucleoside reverse-transcriptase inhibitors (NRTIs), but not highly active antiretroviral therapy (HAART) [45].

3.2.2 Relating Expression Data to Other Biological Information

To gain insight into the biological process and to make new discoveries, the goal is to link gene expression profiles with external information. Some of the possible results that can be obtained by analyzing gene expression data are the ability to predict protein interactions, and their functions [47]. Incorporating expression data with other external information, for example, metabolic pathways of proteins, has been used to predict interacting proteins, protein complexes, and protein function [47]. Genes with similar expression profiles are more likely to encode proteins that interact [48].

3.2.3 Pathway Studio 9.0 Methods

Pathway Studio Software, version 9.0, which is a pathway analysis tool supplied with the RESNET database, harvests the latest information from deposited literature in PubMed and other public sources. The software also uses a number of public and commercial databases, i.e., KEGG (metabolic database; http://www.genome.jp/kegg/), BIND (protein interaction database; http://www.bind.ca), and GO (Gene Ontology) http://www.geneontology.org/. The RESNET product includes a database of relations for mammalians and plants. For this work, we selected direct interactions, shortest path and the common regulators algorithms to build a network among the HPPCR-HIV pathogenesis and other human proteins in a cell. Relationships between HPPCR-HIV pathogenesis and other entities were identified using the following relation type filter parameters: Binding, PromoterBinding, ProtModification, miRNAEffect, Direct regulation and MolTransport. We applied one of the most stringent GeneSpring testing corrections called Bonferroni (Single Step) in this work. In Bonferroni correction, the p-

value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant: i.e.:

$$\text{Corrected P-value} = \text{p-value} * n \text{ (number of genes in test)} < 0.05$$

As a consequence, with 1000 genes at a time, the highest accepted individual p-value is 0.00005, making this correction method very stringent. With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05[49].

3.3 Results and Discussion

Here, we propose a supervised screening framework to select genes from our processed data-set. Genes with a positive differential expression values and p-values less than 0.05 are considered to be up-regulated candidates, while genes with negative values of differential expression and p-values less than 0.05 are down-regulated candidates. The HPPCR-HIV pathogenesis identified in the first assessment (see Table 3.1) were also added when the network was constructed. For testing purposes in the PPIN prediction task, it is therefore common to choose protein pairs uniformly, which has higher connections from the set of protein pairs that are known to interact [50]; thus, we will focus on IL10, CD4, HIVEP2, CCR5 and CXCR4 HPPCR-HIV pathogenesis.

3.3.1 Shortest Path Networks

Biologically, it is of interest to identify the features that contribute the most to the classification of protein pairs. This not only helps reveal relationships between proteins, but also can suggest interactions in the human genome system [51]. We assessed the

shortest path interactions of HPPCR-HIV pathogenesis and other human proteins based on the miRNAEffect relation type. Strikingly, Figure 3.1 shows the extent to which the selected proteins (17 out of 19) have a direct and indirect relationship with their neighbors. Each interaction in the Pathway Studio network is represented by different colors and symbols of arrows and lines.

The IL10, CD4, CCNT1, HIVEP1 and HIVEP2, HPPCR-HIV pathogenesis have many associations with other human proteins in the cell. The proteins called CCR5, CCR3, CD4 and CXCR4 are binding directly to each other. A transcription factor, peroxisome proliferator-activated receptor gamma (PPARG), regulates and changes the localization of IL10 by molecular transport interaction. Interestingly, when the chemokine (C-C motif) ligand 5 (CCL5) proteins induced the expression of CCR5, confocal laser microscopy revealed that CCR5 was colonized with CXCR4 on the cell surface. The promoterBinding of the regulatory factor x 1(RFX1) to CD4 gene shows we can suppress the expression of CD4 by controlling RFX1 in human cell. The networks of some HPPCR-HIV pathogenesis (i.e., OTUD4, AGFG2, TARBP1 and HTATIP2), as shown in Figure 3.1, are controlled by microRNAs; conversely, other HPPCR-HIV pathogenesis (i.e., KAT5, VPRBP, ITIH4, CCNT1 and HTATSF1) are not controlled by microRNAs.

Some entries from the HPPCR-HIV pathogenesis (i.e., HIVEP2, CCNTI and IL10) are more likely associated with other human proteins at least two times in the shortest path networks of HPPCR-HIV pathogenesis and other human proteins PPINs (Figure 3.1).

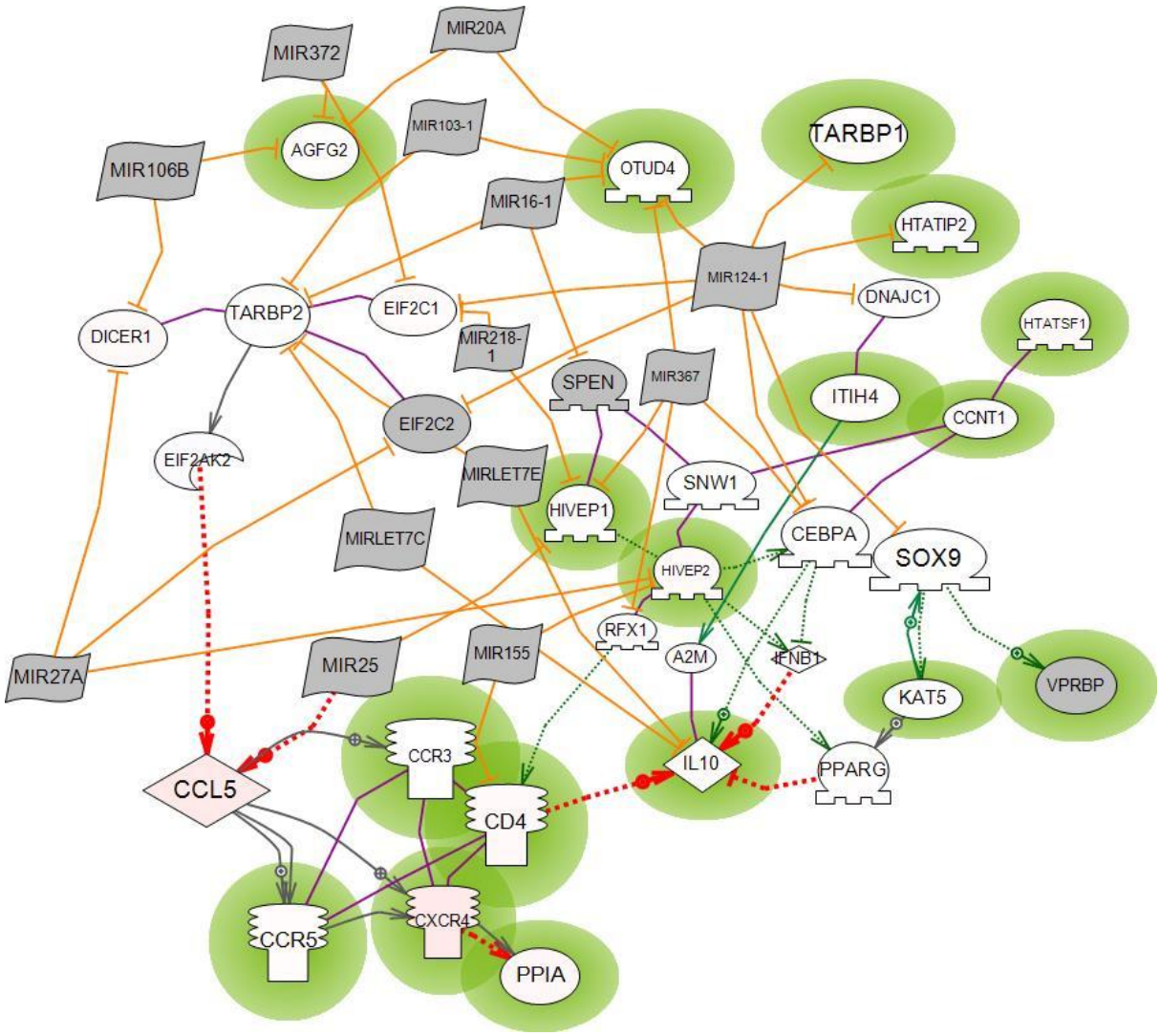


Figure 3.1: The shortest path networks of 19-(HPPCR-HIV pathogenesis) and other human proteins in a cell. A line connecting two nodes indicates the relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green line; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are human proteins.

3.3.2 Common Regulators Networks

Analysis of the common regulator networks of the principal 19-HPPCR-HIV pathogenesis displayed multiple interactions for a subset of 15 human proteins that play critical role in HIV pathogenesis in a human cell. Based on the ProtModification interaction between G protein–coupled receptor kinase 5 (GRK5) or GRK6 and adrenergic beta receptor kinase 1 (ADRBK1) or ADRBK2, we can suppress the HPPCR-HIV pathogenesis CXCR4 and CCR5 as shown in Figure 3.2. As discussed above for the shortest pathway networks, the protein OTUD4 is controlled by MIR142, MIR124-1, MIR367, MIR16-1 and MIR20A microRNAs. The expression of the eminent protein, IL10, is cooperatively activated by the transcription factors CEBPB, STAT4, JUN and IRF3.

CD4 has 13 interactions with other human proteins in the cell. Human proteins ELANE, CTSG and SYK have regulators that change the modification of CD4 in a cell by phosphorylation. Five out of the thirteen interactions are DirectRegulation, which influence CD4 activity by direct physical interaction while the rest are regulators that bind to the promoter of the HPPCR-HIV pathogenesis.

Here, HIVEP2 has no correlation with other human proteins, but IL10 has 10 promoterBinding interactions with other proteins in the human cell. CXCR4 and CCR5 have the most interactions on this network, which means these proteins are regulated by a large number of human proteins in a cell.

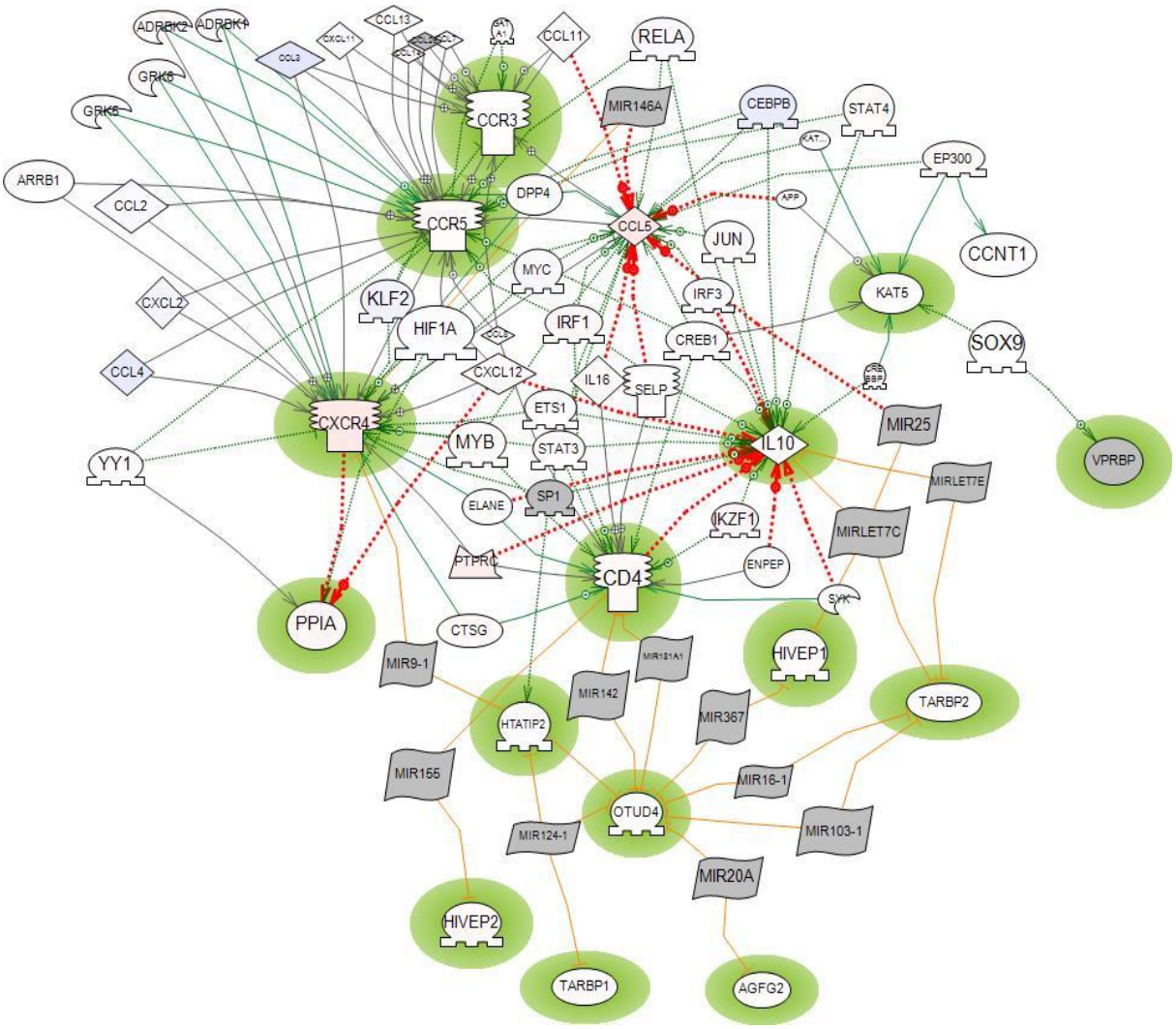


Figure 3.2: The common regulators networks of 19-(HPPCR-HIV pathogenesis) in a cell. A line connecting two nodes indicates the relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green lines; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

Hereafter, we devise our network analysis in four distinct expression sets:

A. **Expression Set One**: HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count decreases and the drug regimen not indicated;
B. **Expression Set Two**: HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count increases and the drug regimen not indicated;
C. **Expression Set Three**: HIV-1 seronegative vs. HIV-1 seropositive expression with unknown CD4 count and the drug regimen is not indicated;
D. **Expression Set Four**: HIV-1 seronegative vs. HIV-1 seropositive expression with unknown CD4 count and drug-naïve.

3.3.3 Direct Interactions Networks for all Expression Sets

Ten out of nineteen, HPPCR-HIV pathogenesis interact via DirectRegulation, Binding, PromoterBinding and MolTransport with other human proteins in the cell as shown in Figure 3.3. For instance, CCR5 has 5 DirectRegulation interactions, out of which 3 are with other HPPCR-HIV pathogenesis (i.e., CCL5 and CXCR4). This revealed that HPPCR-HIV pathogenesis are also capable of suppressing other human proteins that play critical role in HIV pathogenesis (i.e., CCR5). Moreover, CCR3, CD4 and CXCR4 are connected to each other with Binding interactions.

The interactions between IL10 and CEBPB proved that the encoded protein CEBPB is important in the regulation of genes involved in immune and inflammatory responses. Also, Qadri et al. stated CEBPB binds to the IL-1 response element in the IL-6 gene [52]. Besides the Binding connections, CD4 exhibits zero interactions with other human proteins in the cell on this **Expression Set One** direct interaction network.

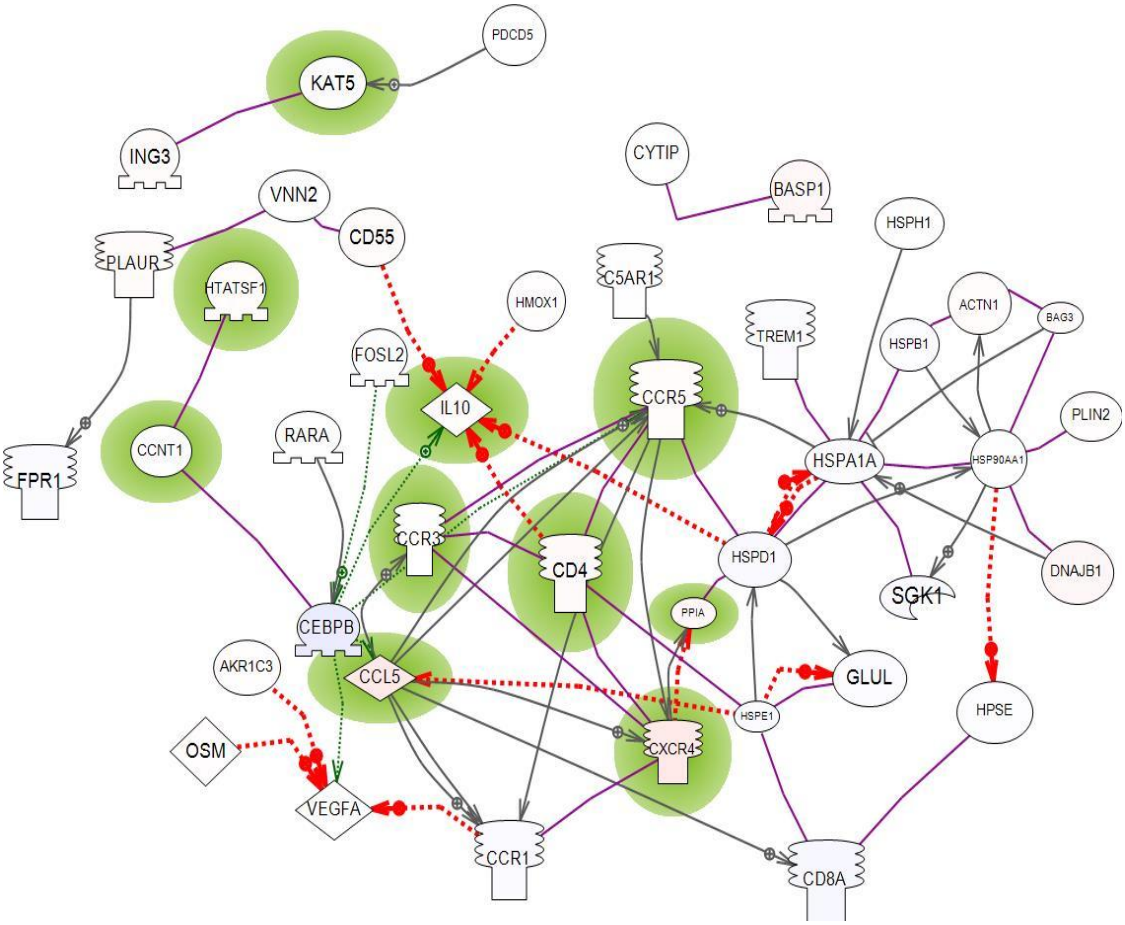


Figure 3.3 HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count decreases and drug regimen not indicated (**Expression Set One**): Direct interaction networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green lines; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

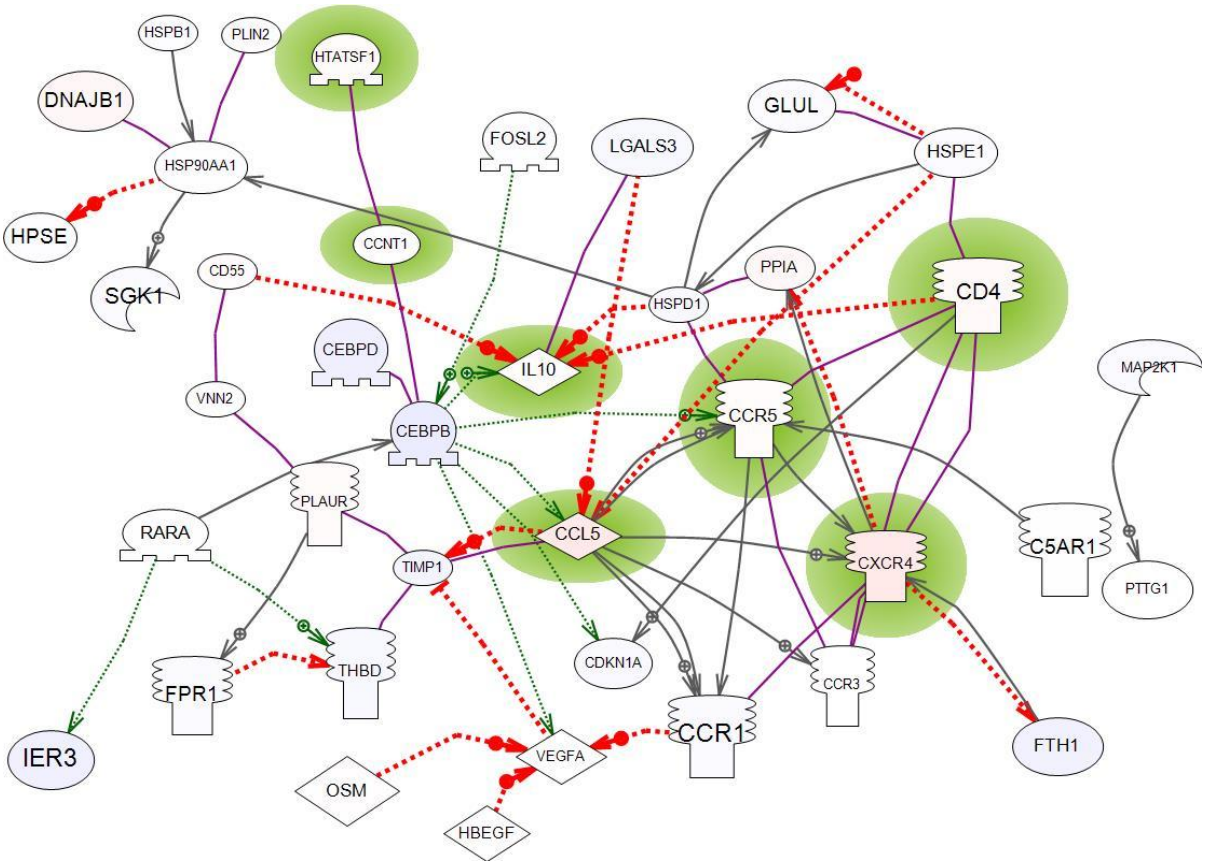


Figure 3.4 HIV-1 seronegative vs. HIV-1seropositive expression when CD4 count increase and drug regimen not indicated **(Expression Set Two):** Direct interaction networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green lines; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

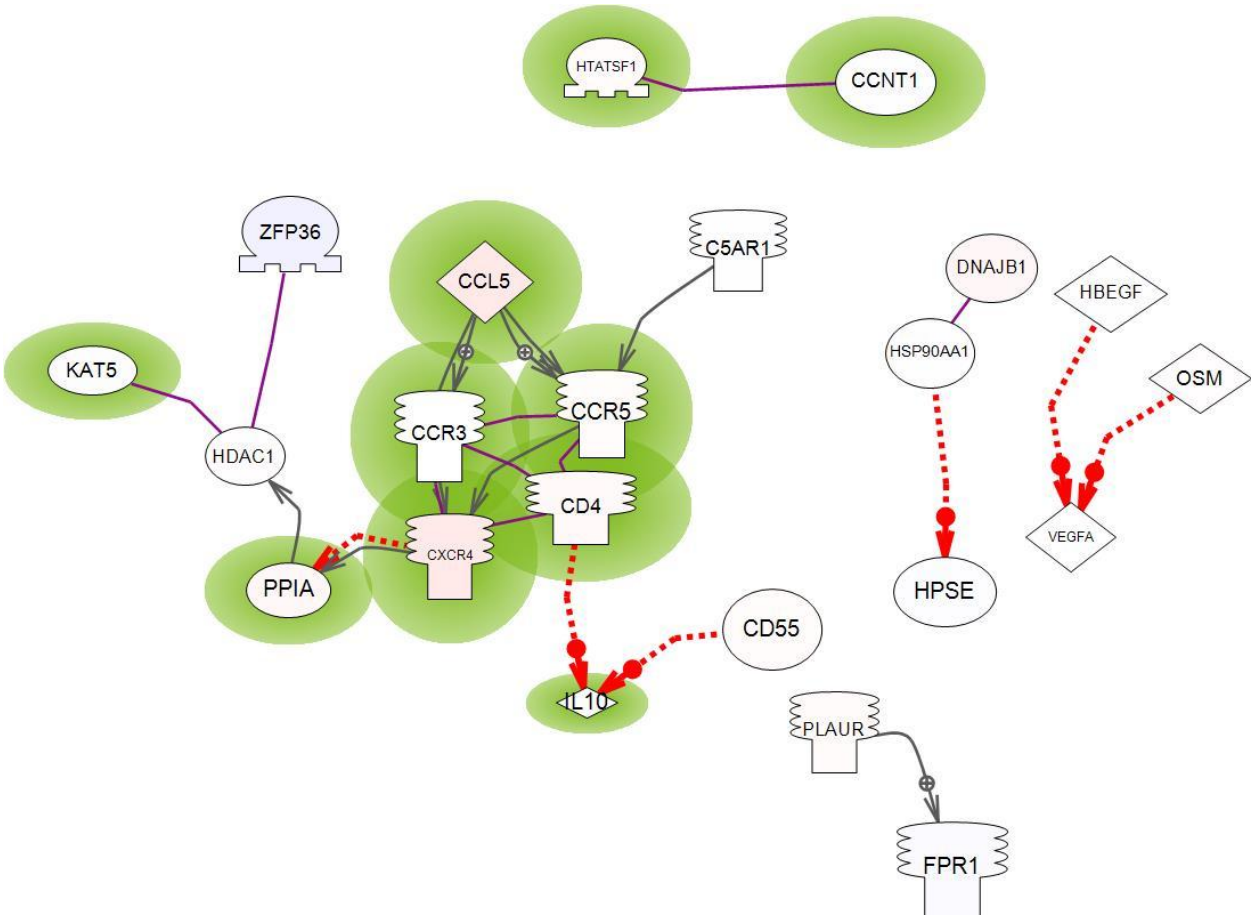


Figure 3.5 HIV-1 seronegative vs. HIV-1seropositive expression when CD4 count is unknown and drug regimen not indicated **(Expression Set Three)**: Direct Interaction networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green lines; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

As was shown above, CCR5, CD4 and CXCR4 proteins also have similar Binding interactions in the HIV-1 seronegative vs. HIV-1 seropositive expressions of

(Figure 3.4). The HPPCR-HIV pathogenesis (i.e., CXCR4) is also mediated by ferritin heavy polypeptide 1(FTH1) chain nuclear translocation in the network as described by Li et al [53]. However, the PromoterBinding interactions of the transcription factor CEBPB with CCL5 and IL10 elucidate that the factor has a negative influence on the expressions of the proteins.

The ligand HPPCR-HIV pathogenesis (i.e., CCL5) has DirectRegulation interactions with the receptor proteins CCR3 and C CR5 (Figure 3.5). In this direct interaction network the HPPCR-HIV pathogenesis (i.e., PPIA), which had not appeared in the previous analysis, has a DirectRegulation interaction with histone deacetylase 1 (HDAC1). In this expression, three direct interaction networks for the interaction of human proteins that play critical role in HIV pathogenesis with other human proteins is mostly insignificant, only 2 of the HPPCR-HIV pathogenesis (i.e., KAT5 and PPIA) have Binding and DirectRegulation connections with histone acetylation and deacetylation (HDAC1) human protein.

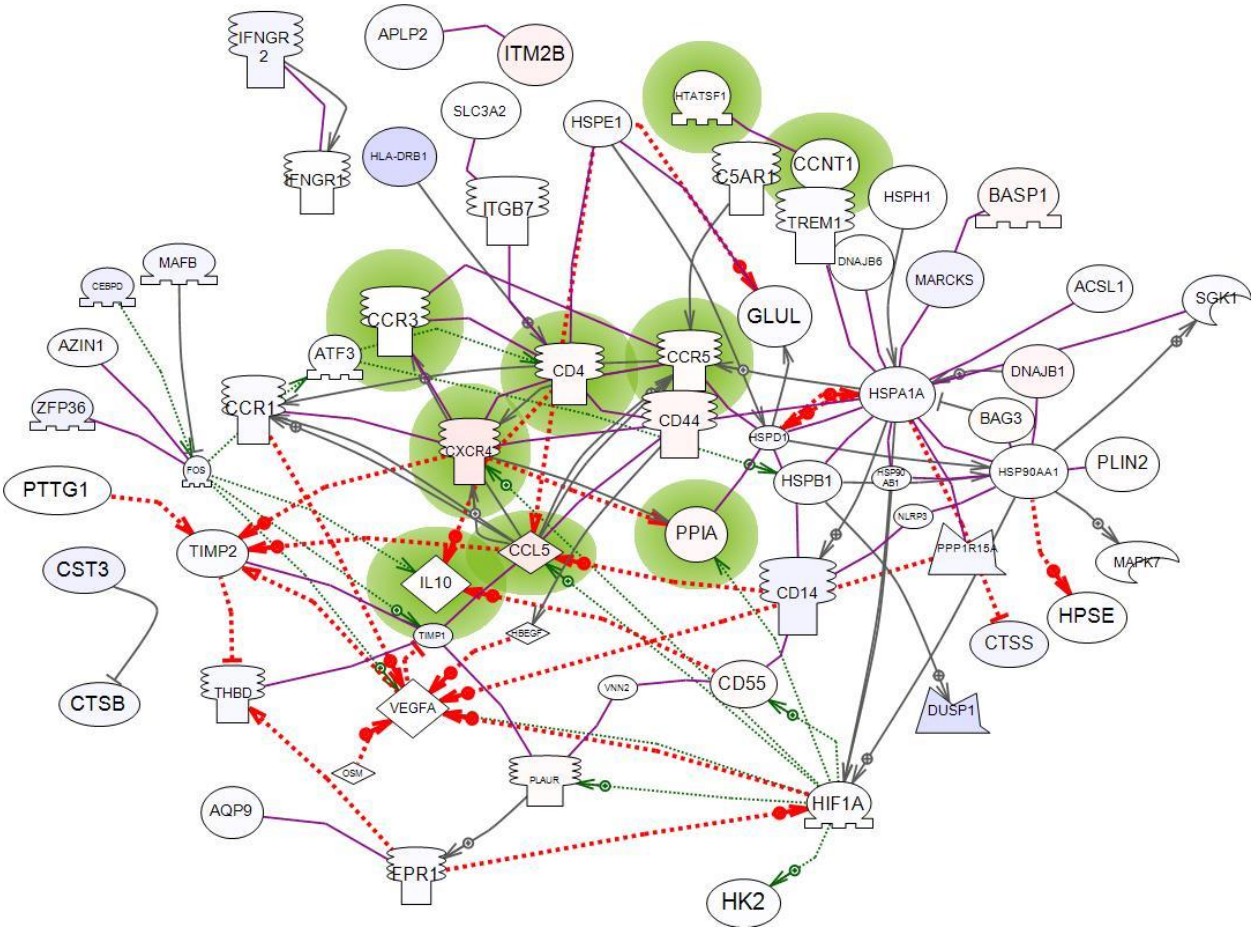


Figure 3.6 HIV-1 seronegative vs. HIV-1seropositive expression when CD4 count is unknown and drug-naïve **(Expression Set Four)**: Direct Interaction Networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. Binding relationships are shown in purple; PromoterBinding relationships are shown in dashed green lines; ProtModification relationships are shown in solid green lines; miRNAEffect relationships are shown in orange; DirectRegulation relationships are shown in black lines; and MolTransport relationships are shown in dashed red lines. The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

On the last expression of HIV-1 seronegative vs. HIV-1 seropositive when the CD4 count is unknown and drug-naïve, the direct interactions network (Figure 3.6) is a significantly more complex network diagram compared to the other direct interaction networks. Only ten out of nineteen, HPPCR-HIV pathogenesis are observed in the direct interaction diagram, although most of them appeared in the previous networks. Here they exhibit somewhat different interactions. The peptidylprolyl isomerase A (PPIA) has a Binding (i.e., a directly physical) interaction with the heat shock 60kDa protein 1 (HSPD1) and PromoterBinding interactions with the hypoxia inducible factor 1 (HIF1A) transcription factor. The direct physical interactions of CCR5 with the heat shock 70kDa protein 1A (HSPA1A) indicates that the expression of CCR5 is controlled by HSPA1A, which has different kinds of interactions with other proteins in the human cell.

The number of interactions on **Expression Set Three** is less compared to the other **Expression Sets**, which indicates the samples under that set is treated by medications that inhibit the replication of the virus in the cell.

3.3.4 Intersection (Shortest Path and Common Regulators) Networks for all Expression Sets

In the intersection of the shortest path and common regulator networks that shows HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count decreases and drug regimen is not indicated. Amazingly, all of the 11 proteins are controlled by microRNAs (Figure 3.7). This means that most microRNAs are involved in suppressing the translation of mRNA of the HPPCR-HIV pathogenesis. For example,

OTUD4 is controlled by MIR101-1, MIR153, MIR103-1, MIR153-1, MIR223, MIR9-1, MIR367 and MIR30B.

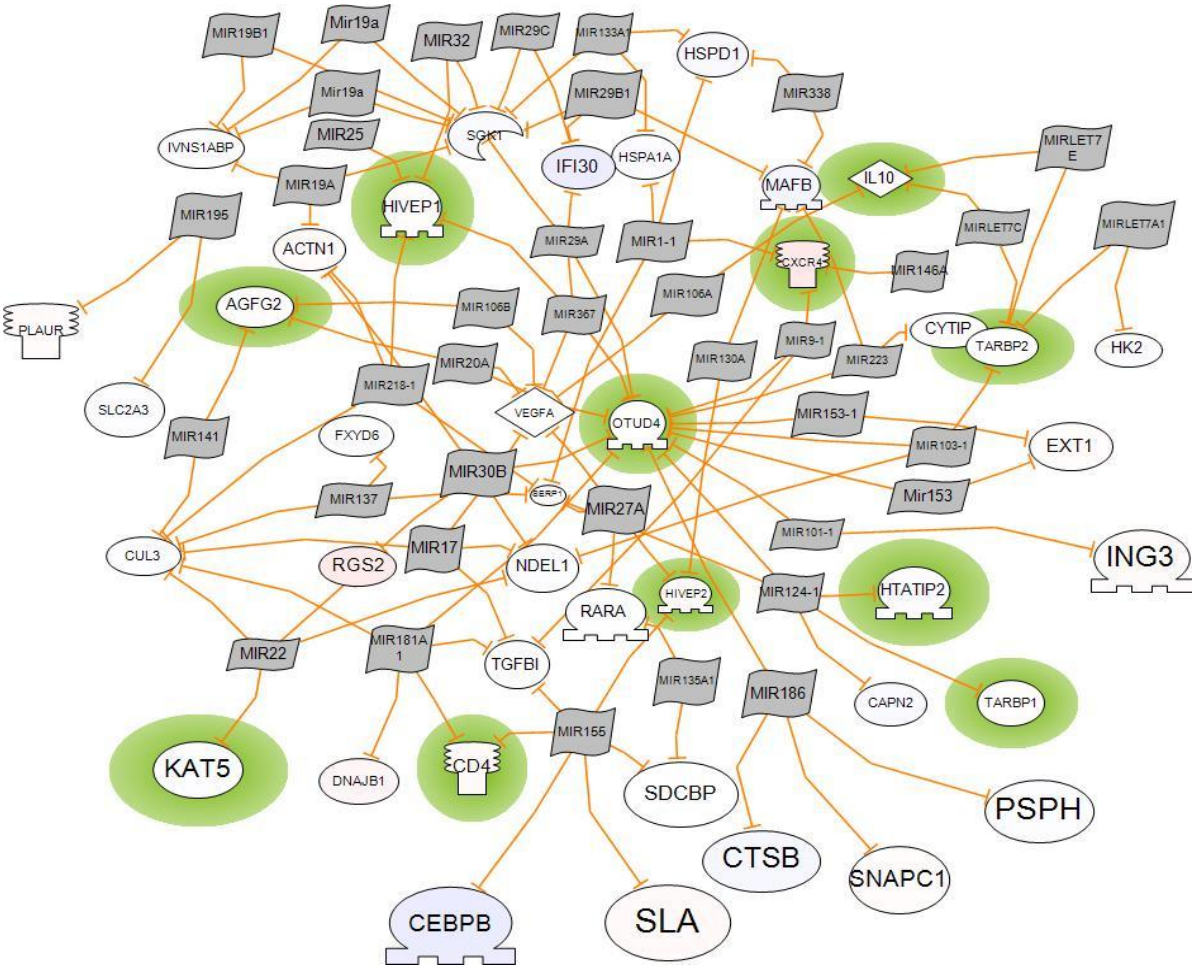


Figure 3.7. HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count decreases and drug regimen is not indicated **(Expression Set One):** Intersection of shortest path and common regulator networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. miRNAEffect relationships are shown in orange; The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

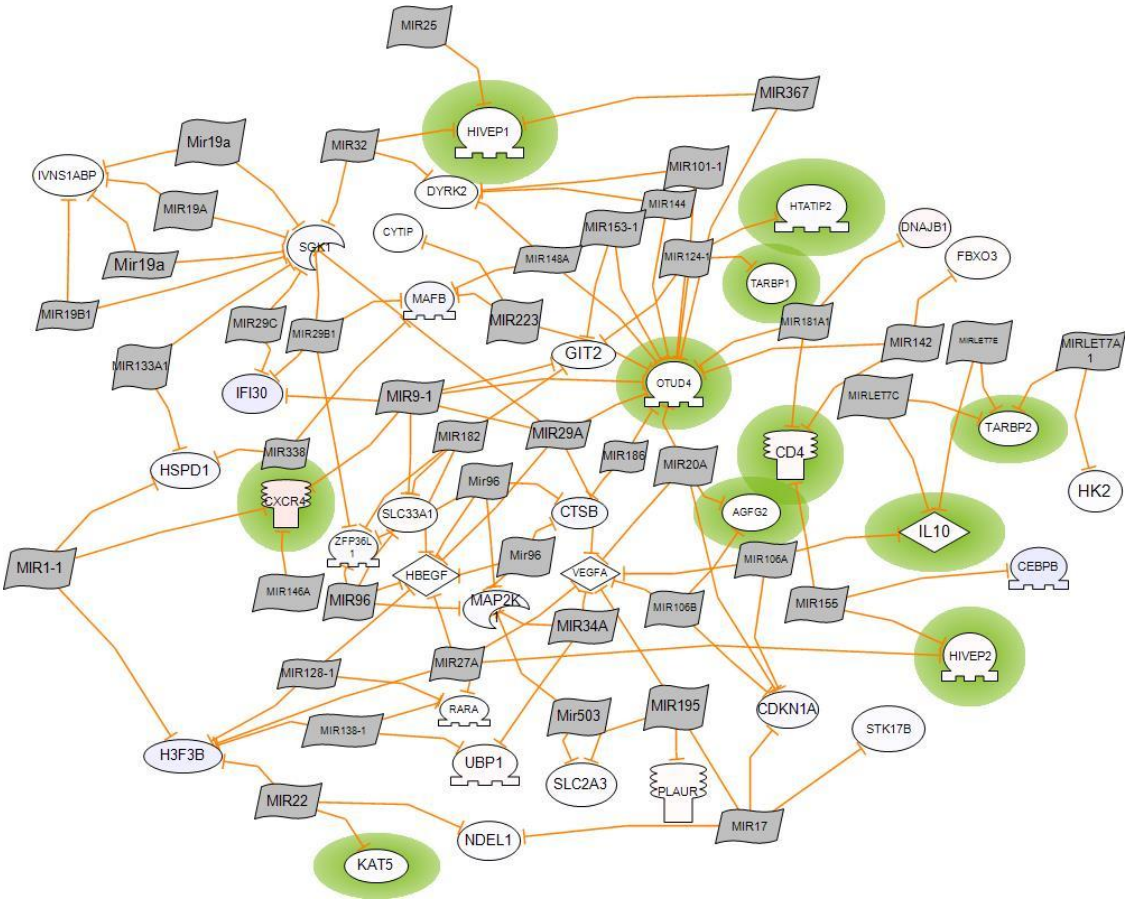


Figure 3.8 HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count increases and drug regimen is not indicated **(Expression Set Two)**: Intersection of Shortest path and common regulator networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. miRNAEffect relationships are shown in orange; The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

The miRNAEffect interactions (i.e., miRNA induced suppression of the gene mRNA) are the most predominant interactions in the intersection of the shortest path and commonRegulators networks of the HIV-1 seronegative vs. HIV-1 seropositive expression when the CD4 count increases and drug regimen is not indicated (Figure 3.8). The translation of almost all human genes that play critical roles in HIV

pathogenesis in the network is controlled by different kinds of microRNAs; for example, we can suppress HIVEP1 in human cells by harnessing the microRNAs (i.e., MIR25, MIR32 and MIR367) from the human immune system.

Like in the previous intersection networks, the networks in Figure 3.9 are also dominated by miRNAEffect interactions that suppress HPPCR-HIV pathogenesis translation.

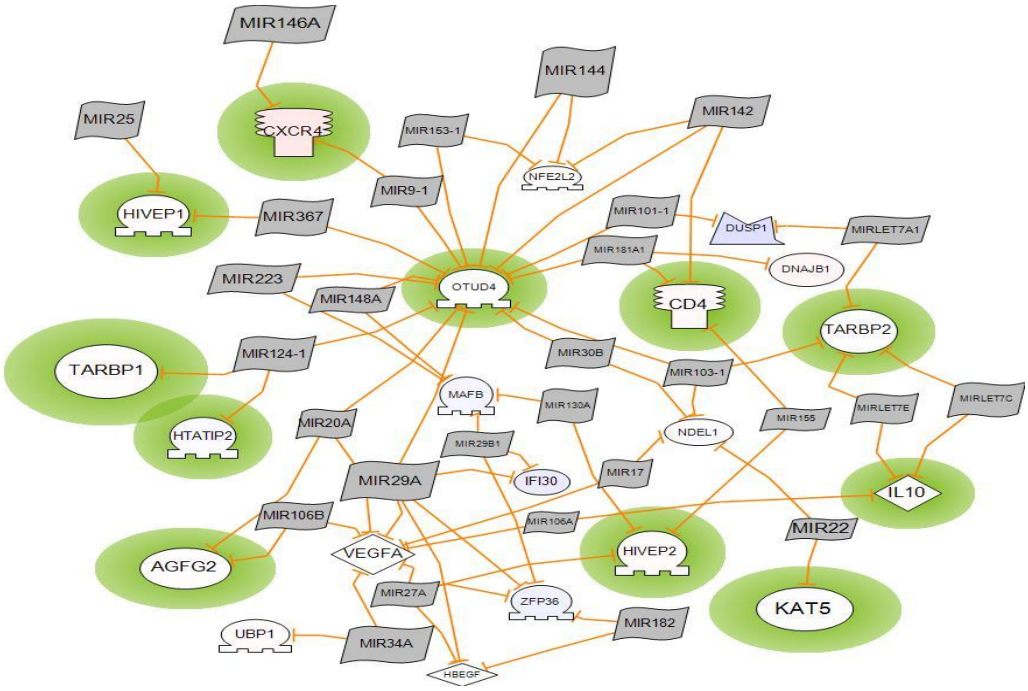


Figure 3.9.HIV-1 seronegative Vs. HIV-1 seropositive expression when CD4 count is unknown and drug regimen is not indicated **(Expression Set Three)**: Intersection of shortest Path and common regulator networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. miRNAEffect relationships are shown in orange; The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

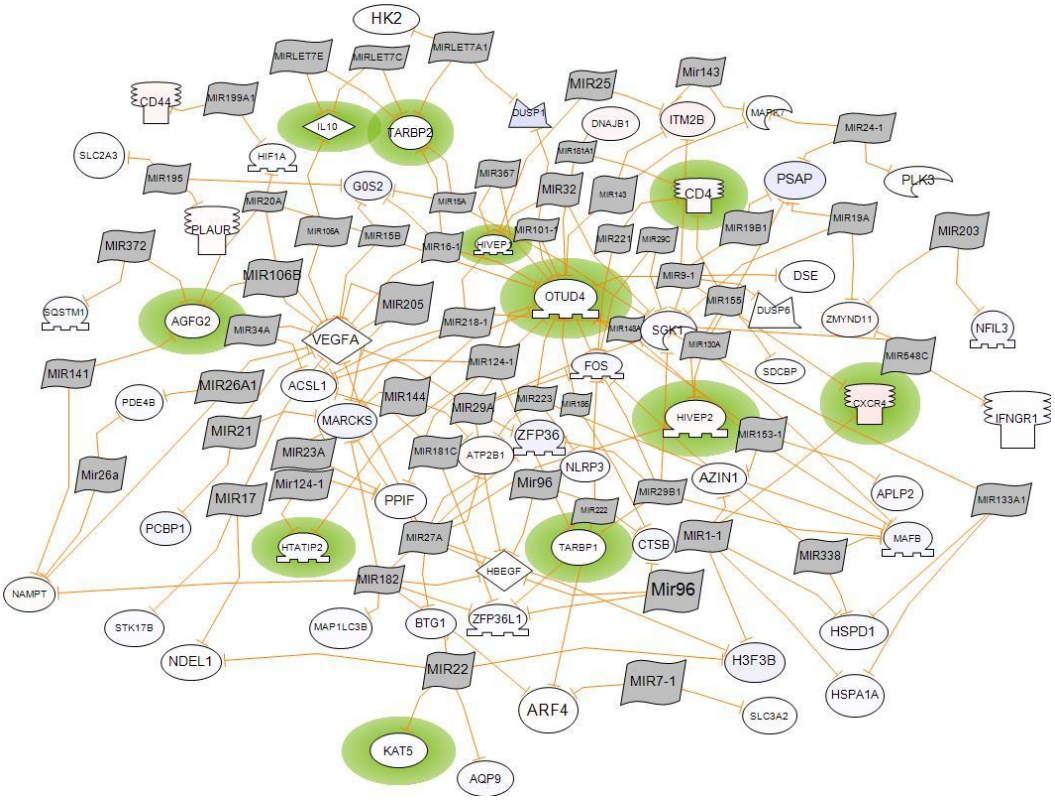


Figure 3.10. HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count is unknown and drug-naïve **(Expression Set Four)**: Intersection of shortest path and common regulators networks. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. miRNAEffect relationships are shown in orange; The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

OTUD4 is regulated by the highest number of microRNAs in the intersection of the shortest path and common regulator networks although it vanished in the direct interactions networks of the expression (Figure 3.10). Repetitively, MIR9-1, MIR367, MIR16-1, MIR148A, MIR144 and MIR218-1 microRNAs are controlling the expression of OTUD4 HPPCR-HIV pathogenesis.

Next to OTUD4, TARBP2 is regulated by the highest number of microRNAs. However, the networks are the same for all four different expressions (Table 3.2), the interactions between HPPCR-HIV pathogenesis and other human proteins are best seen in the direct interaction networks of **Expression Set One** and **Two**. The numbers of microRNAs that control HPPCR-HIV pathogenesis are similar in most cases, except HIVEP2, which is controlled by four miRNAs in **Expression Set One** and **Four**.

Table 3.2 The availability of HPPCR-HIV pathogenesis in each network for each **Expression Sets** and the number of microRNAs interactions with each HPPCR-HIV pathogenesis.

	Expression Set One			Expression Set Two			Expression SetThree			Expression Set Four		
HIV-Proteins	Direct Interaction Networks	Intersection Networks	Number of Associated microRNAs	Direct Interaction Networks	Intersection Networks	Number of Associated microRNAs	Direct Interaction Networks	Intersection Networks	Direct Interaction Networks	Direct Interactions Networks	Intersection Networks	Direct Interaction Networks
OTUD4	×	√	13	×	√	13	×	√	14	×	√	14
CCR5	√	×		√	×		√	×		√	×	
CCL5	√	×		√	×		√	×		√	×	
CXCR4	√	√	3	√	√	3	√	√	2	√	√	2
CD4	√	√	2	√	√	3	√	√	3	√	√	3
IL10	√	√	3	√	√	3	√	√	3	√	√	3
KAT5	√	√	1	×	√	1	√	√	1	×	√	1
PPIA	√	×		×	×		√	×		√	×	
CCR3	√	×		×	×		√	×		√	×	
CCNT1	√	×		√	×		√	×		√	×	
HIVEP1	×	√	4	×	√	3	×	√	2	×	√	4
HTATIP2	×	√	1	×	√	1	×	√	1	×	√	2
HTATSF1	√	×		√	×		√	×		√	×	
HIVEP2	×	√	3	×	√	2	×	√	3	×	√	3
VPRBP	×	×		×	×		×	×		×	×	
ITIH4	×	×		×	×		×	×		×	×	
TARBP2	×	√	4	×	√	3	×	√	4	×	√	5
TARBP1	×	√	1	×	√	1	×	√	1	×	√	1
AGFG2	×	√	3	×	√	2	×	√	2	×	√	4

Keys: (√ means present; × means absent)

Expression Set One: HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count decreases and drug regimen is not indicated; **Expression Set two**: HIV-1 seronegative vs. HIV-1 seropositive expression when CD4 count increases and drug regimen is not indicated; **Expression Set Three**: HIV-1 seronegative vs. HIV-1seropositive expression with unknown CD4 count and drug regimen is not indicated; **Expression Set Four**: HIV-1 seronegative vs. HIV-1 seropositive expression with unknown CD4 count and drug-naïve.

The combined pathway of all four expression sets intersection networks gives us integrated human proteins that play critical role in HIV pathogenesis /other human proteins/microRNAs network. In this integrated network diagram, we have determined that most human genes that encode HPPCR-HIV pathogenesis are suppressed by some microRNAs.

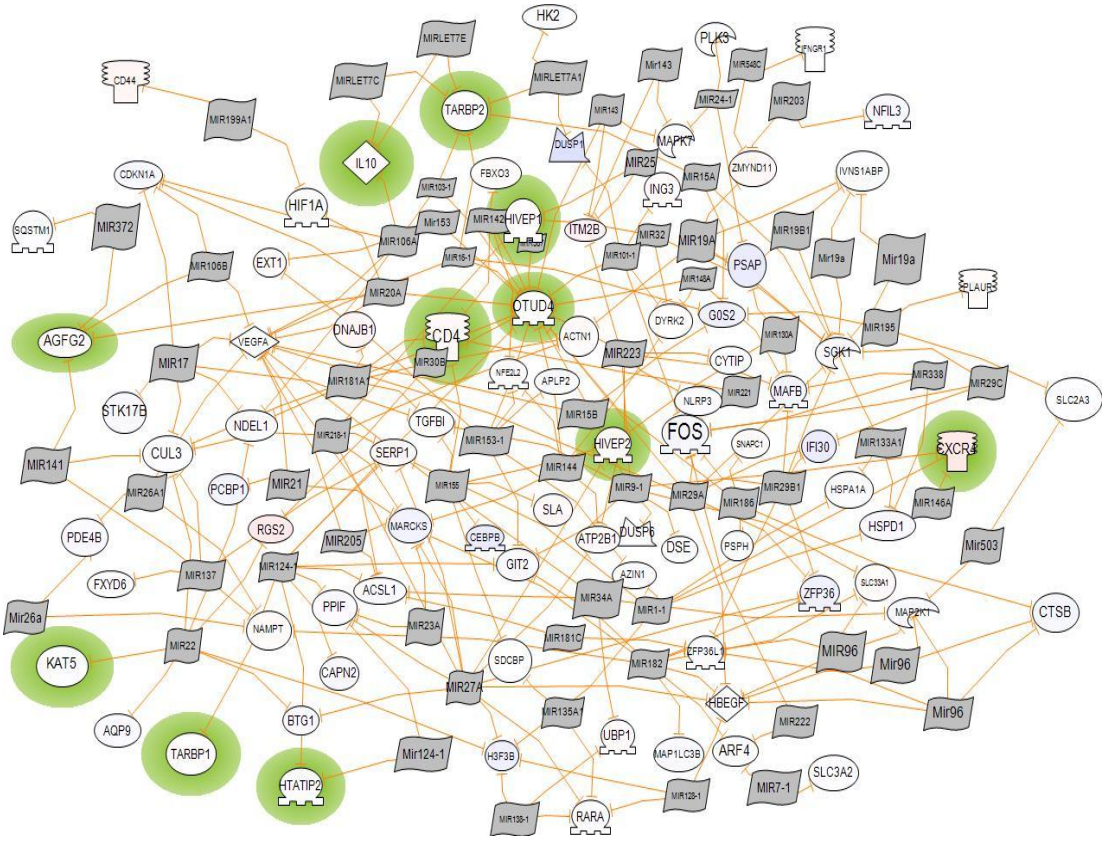


Figure 3.11 An Integrated HPPCR-HIV pathogenesis /other human Proteins/ microRNAs interaction network. A line connecting two nodes indicates relationship type between HPPCR-HIV pathogenesis and other human proteins or microRNAs. miRNAEffect relationships are shown in orange; The nodes highlighted by green bubbles are HPPCR-HIV pathogenesis; nodes that are in gray bubbles are microRNAs; the remaining nodes are other human proteins.

Table 3.3 Pathway Studio Gene Ontology enrichment analysis of HPPCR-HIV pathogenesis-microRNAs interaction Network.

Name	Overlap	% Overlap	Overlapping Entities	p-value	Data Source
RNA Gene Silencing	4	3	MIRLET7A1,MIRLET7C,TARBP2,MIRLET7E	0.005	CPP
Mitochondrial Protein Transport	4	3	HSPA1A,HSPD1,DNAJB1,PPIF	0.005	CPP
Focal Adhesion Regulation	7	2	HBEGF,VEGFA,MAP2K1,DUSP1,ACTN1,DUSP6,SDCBP	0.010	CSP
NGFR -> AP-1/CEBPB/CREB/ELK-SRF/TP53 signaling	3	7	FOS,MAP2K1,CEBPB	0.009	RSP
EctodysplasinR -> AP-1 signaling	2	10	FOS,MAP2K1	0.021	RSP
IL6R -> CEBP/ELK-SRF signaling	2	8	MAP2K1,CEBPB	0.032	RSP
AdenosineR -> AP-1 signaling	2	8	FOS,MAP2K1	0.035	RSP
IL5R -> SOX4 signaling	1	25	SDCBP	0.047	RSP
response to stress	9	3	FOS,HSPA1A,SGK1,IL10,CTSB,SQSTM1,RARA,DNAJB1	0.000	BP
negative regulation of apoptosis	9	3	VEGFA,HSPA1A,CDKN1A,HSPD1,SGK1,CD44,PLAUR,S	0.000	BP
cell migration	7	4	HBEGF,VEGFA,CD44,CXCR4,CUL3,BTG1,NDEL1	0.000	BP
response to organic cyclic compound	8	3	FOS,CDKN1A,HSPD1,CD44,ACSL1,NAMPT,DUSP6,CTS	0.000	BP
MyD88-dependent toll-like RSP	5	6	FOS,MAP2K1,HSPD1,MAPK7,DUSP6	0.000	BP
anti-apoptosis	7	2	VEGFA,HSPA1A,CEBPB,HSPD1,IL10,SQSTM1,HTATIP2	0.000	BP
response to organic substance	6	3	CDKN1A,DUSP1,HSPD1,AQP9,ACSL1,IL10	0.000	BP
regulation of apoptosis	7	2	HIF1A,DUSP1,ACTN1,CTSB,BTG1,HTATIP2,PPIF	0.000	BP
phosphorylation	9	1	MAP2K1,CDKN1A,SGK1,MAPK7,PLAUR,HK2,STK17B,(	0.000	BP
stress-activated MAPK cascade	4	7	FOS,MAP2K1,MAPK7,DUSP6	0.000	BP
protein heterodimerization activity	9	2	VEGFA,FOS,CAPN2,HIF1A,CEBPB,HSPD1,RARA,SDCBI	0.000	MF
protein homodimerization activity	10	1	VEGFA,CEBPB,ACTN1,CD4,NAMPT,SQSTM1,SDCBP,EX	0.000	MF
protein complex binding	6	2	HIF1A,CDKN1A,HSPD1,CTSB,KAT5,NDEL1	0.000	MF
kinase activity	10	1	MAP2K1,CDKN1A,SGK1,MAPK7,PLAUR,HK2,STK17B,I	0.000	MF
glucose binding	2	15	SLC2A3,HK2	0.001	MF
AU-rich element binding	2	14	ZFP36,ZFP36L1	0.001	MF
MAP kinase tyrosine-serine-threonine phosphatase activity	2	13	DUSP1,DUSP6	0.001	MF
cyclin binding	2	12	CDKN1A,CUL3	0.001	MF
cell surface binding	2	11	VEGFA,HSPD1	0.001	MF

CPP	Cell Process Pathways
CSP	Cell Signaling Pathways
RSP	Receptor Signaling Pathways
BP	biological_process
MF	molecular_function

3.3.5 Pathway Studio Gene Ontology Enrichment Analysis

It can clearly be seen that four of the expressed genes from the cell process pathway are associated with RNA gene silencing and another four are also found for mitochondrial protein transport. The most interesting group attribute in the receptor signaling pathways is that of the interleukin-5 receptor (IL5R), which belongs to the type I cytokine receptor family and is a heterodimer composed of two polypeptide chains, exclusively expressed by the transcriptional factor SOX-4. SOX4 is expressed in lymphocytes (B and T) and is required for B lymphocyte development.

The size of the groups that are involved in biological processes are much larger than that expected by chance for this process, meaning that they are over-represented. The most significantly upregulated genes (i.e., p-value < 0.01), are associated with anti-apoptosis and response to stress. Particularly, a gene called vascular endothelial growth factor-A(VEGF-A),which is under the control of many microRNAs (Figure 3.11), has various effects, including promoting cell migration and inhibiting apoptosis as it is shown in (Table 3.3).

The highly enriched network (i.e., CD4, KAT5) from the molecular function category carry out the protein is complex binding functions in a cell. This means that they interact selectively and non-covalently in the cell with any protein complex (a complex of two or more proteins that may include other non-protein molecules). Even though it has seemingly few occurrences, kinase activity is one of the functions carried out by some genes that encode for network; however, the expression of the human genes that play critical role in HIV pathogenesis can be suppressed by some microRNAs.

3.4 Conclusion:

Computational methods can be very effective in assisting experimental efforts to show interacting protein pairs within a single organism. This study applied the Pathway Studio software to build networks integrating human proteins that play critical role in HIV pathogenesis and other human proteins that interact with each other, as well as networks involving miRNAs that target mRNAs of genes encoding network. Features derived from multiple genomic and functional data sources, coupled with exploiting our knowledge of the human proteins interactome, were integrated in a supervised learning framework. Hence, we constructed the human protein-protein pathways using the microarray data of functional genomic relationships in HIV-1 disease to expedite the elucidation of the important mechanisms of HIV-human cell interactions and their implications.

In the final PPI networks of our analysis, we generated a new hypothesis based on the commonality of the selected proteins in a cell. Most of the human proteins we used for the analysis are regulatory proteins that drastically enhance the efficiency of HIV virus; this in turn allows us to be able to understand the association of each protein with other human proteins that have fewer roles in HIV pathogenesis and microRNAs. Thus, human proteins which have critical role in HIV pathogenesis interacted with other common human proteins and they are regulated by common transcription factors. Finally, the important practical aspect of this study offers many options for suppressing the expression of the human-genes that play critical role in HIV pathogenesis, and thus could of interest in developing new anti-HIV drugs.

Chapter 4

CLOSING REMARKS

An ample description on protein-protein interactions would involve the structure of the proteins and everything that have associations with them. Since proteins play great roles in carrying out biochemical functions in a living cell, many studies have been undertaken to understand how they are localized, how they are regulated and how they are crystallized. But, we should keep in mind that we are still in the process of qualitatively cataloging protein-protein interactions and paying too much attention to the quantitative and dynamic aspects may be premature for many cases [54].

In this report we have tried to analyze one seemingly small but very important issue, the characteristics of interfacial water molecules at protein-protein interfaces. This will help us, as we begin to gain an understanding of the interactions of polar and hydrophobic biologically relevant proteins. The availability of information about interfacial water molecules could assist our understanding and localizations on the type of each residue when we are designing crystallization experiments.

In addition to the studies on the role of water molecules in the binding sites of protein-protein complexes, network analyses of PPIs are undoubtedly powerful as they give specific functional implications of an interaction. To date, we understand only 10% of all human protein-protein interactions [55] and some recent studies estimate that we have identified only 50% of all yeast interactions; hence, in order to know the dynamics and kinetics of protein complexes, we first must explore their interactions via bioinformatics tools. The study on HPPCR-HIV pathogenesis and other human proteins

pathway needs lots of time and huge dataset to be able to address each and every mechanism of the interactions, but in this small dataset and short period of time we are able to identify the regulatory proteins and microRNAs that drastically enhance the replication of HIV in human cell.

Citations:

1. Bahadur, R et al. **(2004)** A Dissection of Specific and Non-specific Protein-Protein Interfaces. J.Mol.Biol.336,943-955
2. Zhanhua, C et al. **(2005)** Protein subunit interfaces: heterodimers versus homodimers. Bioinformation1 (2): 28–39.
3. Nooren, I et al. **(2003)** Transient protein–protein interactions. EMBO J., 22, 3486–3492
4. Bahadur, R et al. **(2007)** The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. Cell. Mol. Life Sci.(65) 1059 -1072
5. Ofran, Y and Rost, B et al. **(2003)** Analysing six types of protein-protein interfaces. J.Mol.Bio.10; 325(2):377-87.
6. Jansen, R., et al. **(2002)** Relating whole-genome expression data withprotein–protein interactions. Genome Res. 12:37–46
7. Tirosh,I and Barkai,N **(2005)** Inferring regulatory mechanisms from patterns of evolutionary divergence. Mol Syst Biol; 7: 530.
8. Zanivan et al. **(2007)**. A new computational approach to analyze human protein complexes and predict novel protein interactions. Genome Biol.; 8(12): R256.

9. Zhu, G et al. **(2006)** CD147 overexpression on synoviocytes in rheumatoid arthritis enhances matrix metalloproteinase production and invasiveness of synoviocytes. Arthritis Res Ther 8(2): R44.

10. Ding, X et al. **(2009)** A Rice Kinase-Protein Interaction Map. Plant Physiology vol. 149 no. 3 1478-1492

11. Zacharias, M et al. **(2010)** Protein-protein Complexes: Analysis, Modelling and Drug Design. J Med Chem 51:3,499-3,506.

12. Oshima K, Ishii Y, Kakizawa S, Sugawara K, Neriya Y, et al. **(2011)** Dramatic Transcriptional Changes in an Intracellular Parasite Enable Host Switching between Plant and Insect. PLoS ONE 6(8): e23242.

13. Zafra-Ruano, A et al. **(2012)** Interfacial water molecules in SH3 interactions: Getting the full picture on polyproline recognition by protein-protein interaction domains. PMID: 22584053

14. Teyra J, Pisabarro MT **(2007)** Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description. Proteins Struct Funct Bioinf 67: 1087–1095.

15. Sonavane and Chakrabarti, **(2008)** Cavities and Atomic Packing in Protein Structures and Interfaces. PLoS Comput Biol 4(9): e1000188.

16. Li et al. **(2012)** Progressive dry-core-wet-rim hydration trend in a nested-ring topology of protein binding interfaces. BMC Bioinformatics 13: 51.

17. Amadasi, A et al. **(2006)** Mapping the energetics of water–protein and water–ligand interactions with the “natural” HINT force field: predictive tools for characterizing the roles of water in biomolecules. J Mol Biol 358: 289–309.

18. Ahmed, M et al. **(2011)** Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. PLoS ONE 6(9): e24712. doi:10.1371/journal.pone.0024712

19. Tang et. al. **(2006)** A simple and reliable approach to docking protein–protein complexes from very sparse NOE-derived intermolecular distance restraints.Journal of Biomolecular NMR (2006) 36: 37–44

20. Keskin, O et al. **(2008)** Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact? Chem. Rev., 108, 1225−1244.

21. Larsen, T et al. **(1998)** Morphology of protein-protein interfaces. Structure.15; 6(4):421-7.

22. Hakes,L et al. **(2008)** Protein Interactions from Complexes: A Structural Perspective. Comp Funct Genomics: 49356.

23. Xia, K et al. **(2006)** Identification of the roliferation/Differentiation Switch in the Cellular Network of Multicellular Organisms. PLoS Comput Biol 2(11): e145.

24. Joachimiak,L et al. **(2006)** Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J Mol Biol 361: 195–208.

25. Kellogg, G et al. **(2000)** Hydrophobicity: Is $LogP_{o/w}$ more than the sum of its parts? Eur J Med Chem 35: 651–661.

26. Burnett, J, et al. **(2000)** Computational methodology for estimating changes in free energies of biomolecular association upon mutation. The importance of bound water in dimer-tetramer assembly for beta 37 mutant hemoglobins. Biochemistry 39: 1622–1633

27. Fornabaio, M et al. **(2004)** Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. J Med Chem 47: 4507–4516.

28. Kellogg GE, Chen D **(2004)** The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. Chem Biodiver 1: 98–105.

29. Amadasi A, Surface JA, Spyrakis F, Cozzini P, Mozzarelli A, et al. **(2008)** Robust classification of “Relevant” water molecules in putative protein binding sites. J Med Chem 51: 1063–1067.

30. Levitt, M. **(1983)** Molecular dynamics of native protein. I. Computer simulation of trajectories. J. Mol. Biol., 168, 595-617.

31. Levitt, M et al. **(1988)** Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol. 201, 751-754.

32. Spyrakis, F et al. **(2007)** Energetics of the protein-DNA-water interaction. BMC Struct. Biol.7, 4.

33. Marabotti, A et al. **(2008)** Energy-based prediction of amino acid-nucleotide base recognition. Journal of Computational Chemistry 29, 1955-1969.

34. Cozzini, Pet al. **(2004)** Free energy of ligand binding to protein: evaluation of the contribution of water molecules by computational methods. Curr. Med. Chem.11, 3093-3118.

35. Burnett, J. C.et al. **(2001)** Computationally accessible method for estimating free energy changes resulting from site-specific mutations of biomolecules: Systematic model building and structural/hydropathic analysis of deoxy and oxy hemoglobins. Proteins:Structure, Function, and Genetics 42, 355-377.

36. The R Project for Statistical Computing. Vienna, Austria: http://www.R-project.org.

37. Albert, R and Barabasi, A **(2002)** Statistical mechanics of complex networks. Reviews of Modern Physics, Volume 74

38. Han, J et al. **(2004)** Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430, 88-93

39. Jäger, S et al. **(2011)** Purification and characterization of HIV–human protein complexes. Elsevier Inc. Methods 53 pp13–19.

40. Jäger, S et al. **(2011)** Global landscape of HIV–human protein complexes. Nature 1, Vol 0.

41. Chen, K et.al **(2012)** Associations between HIV and Human Pathways Revealed by Protein-Protein Interactions and Correlated Gene Expression Profiles PLoS ONE Volume 7.

42. Henschel A, Kim WK, Schroeder M **(2006)** Equivalent binding sites reveal convergently evolved interaction motifs. Bioinformatics 22: 550–555.

43. Tastan, O et al. **(2009)** Prediction Of Interactions Between HIV-1 AND Human Proteins By Information Integration. Pacific Symposium on Biocomputing 14:516-527.

44. Madoz-Gurpide, J et al. **(2003)** Molecular analysis of cancer using DNA and protein microarrays. Adv. Exp. Med.Biol., 532,51-58

45. Ockenhouse, C et al. **(2005)** Functional Genomic Relationships in HIV-11 Disease Revealed by Gene-Expression Profiling of Primary Human Peripheral Blood Mononuclear Cells. JID:191

46. Babu, M et al **(2004)** . Introduction to microarray data analysis - M. Madan Babu* in Computational Genomics (Ed: R. Grant), Horizon Press, U.K.

47. Ge, H, et al. **(2001)** Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet; 29:482–486.

48. Gerstein, M **(2000)** Integrative database analysis in structural genomics 2000 Nature America Inc.• http://structbio.nature.com

49. Dudoit, S. et al. **(2000)** Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12(2002), 111-139

50. A. Ben-Hur and W.S. Noble, BMC Bioinformatics 7 Suppl 1, S2 **(2006).**

51. Liang, H et al. (2007) MicroRNA regulation of human protein–protein interaction network. RNA.13 (9): 1402–1408.

52. Qadri, I et al. **(2012)** Increased phosphoenolpyruvate carboxykinase gene expression and steatosis during hepatitis C virus subgenome replication: role of nonstructural component 5A and CCAAT/enhancer-binding protein β. J Biol Chem.; 287(44):37340-51

53. Li et al. **(2006)** 3'-Poly (A) tail enhances siRNA activity against exogenous reporter genes in MCF-7 cells . J RNAi Gene Silenc ,2(2), 195-204

54. Hua, W. **(2013)** Journal of Computational Biology.20 (4): 344-358.

55. Hart et al. **(2006)** How complete are current yeast and human protein-interaction networks? Genome Biology*, **7**:120*

Appendix I: List of Homodimeric Protein Complexes examined in the study

PDB ID	Chain ID	Chain Length
1H0H	A/B	977
1H54	A/B	754
1EX0	A/B	731
1H41	A/B	708
1DJX	A/B	624
1AOR	A/B	605
1G8M	A/B	593
1B3U	A/B	588
1F0X	A/B	571
1AMU	A/B	563
1ASO	A/B	552
1GNL	A/B	544
1HDH	A/B	536
1F0L	A/B	535
1AUI	A/B	521
1GOI	A/B	499
1E8C	A/B	498
1DDZ	A/B	496
1FEC	A/B	490
1E5X	A/B	486
1DPG	A/B	485
1DNP	A/B	471
1H80	A/B	464
1F60	A/B	458
1GG4	A/B	452
1HEI	A/B	451
1B8A	A/B	438
1H3F	A/B	432
1HQS	A/B	423
1BK5	A/B	422
1EJD	A/B	419
1GIQ	A/B	413
1AJS	A/B	412
1GWI	A/B	411
1DKL	A/B	410
1CQX	A/B	403
1AZT	A/B	402
1FP3	A/B	402
1CHM	A/B	401
1FC4	A/B	401
1AXK	A/B	394
1DQS	A/B	393
1G4M	A/B	393
1CI9	A/B	392
1EI1	A/B	391
1ELU	A/B	390
1FNN	A/B	389
1GDE	A/B	389
1B5P	A/B	385
1EG5	A/B	384
1AJ8	A/B	371
1FN9	A/B	365
1H7S	A/B	365
1F0K	A/B	364
1GU7	A/B	364
1CZF	A/B	362
1BJN	A/B	360
1DOS	A/B	358
1EBF	A/B	358
1C1D	A/B	355
1DYS	A/B	348
1EK6	A/B	348
1GXR	A/B	337
1GXM	A/B	332
1E2K	A/B	331
12AS	A/B	330
1DPJ	A/B	329
1FVR	A/B	327
1GVE	A/B	327
1BLX	A/B	326
1BSL	A/B	324
1F06	A/B	320
1DKU	A/B	317
1DL5	A/B	317
1EFV	A/B	315
1E19.	A/B	314
1H4R	A/B	314
1DMH	A/B	311
1EUD	A/B	311
1F0C	A/B	305
1H1N	A/B	305
1F0Y	A/B	302
1A4I	A/B	301
1DLE	A/B	298

1H3I	A/B	293
1BF6	A/B	291
1E5R	A/B	290
1DBQ	A/B	289
1GUD	A/B	288
1E9G	A/B	286
1GL4	A/B	285
1DQZ	A/B	280
1BKP	A/B	278
1A8U	A/B	277
1DEU	A/B	277
1EKQ	A/B	272
1CP2	A/B	269
1AD1	A/B	266
1EE8	A/B	266
1B9M	A/B	265
1CY9	A/B	264
1DJ0	A/B	264
1GEE	A/B	261
1H32	A/B	261
1G60	A/B	260
1GK9	A/B	260
1E42.	A/B	258
1FJH	A/B	257
1H0B	A/B	256
1G0H	A/B	252
1ABR	A/B	251
1E2W	A/B	251
1F75	A/B	249
1B12	A/B	248

1GEQ	A/B	248
1GV3	A/B	248
1B5E	A/B	246
1H7E	A/B	245
1AGJ	A/B	242
1DEK	A/B	241
1F5V	A/B	240
1B5F	A/B	239
1HW1	A/B	239
1CQ3	A/B	233
1A7T	A/B	232
1FJ2	A/B	232
1DQN	A/B	230
1EKE	A/B	230
1FL1	A/B	230
1GWC	A/B	230
1EZI	A/B	228
1G61	A/B	228
1GXY	A/B	226
1AVW	A/B	223
1EQ9	A/B	222
1EYQ	A/B	222
1EUV	A/B	221
1GQP	A/B	221
1AB8	A/B	220
1AUO	A/B	218
1G57	A/B	217
1EEJ	A/B	216
1A04	A/B	215
1BQU	A/B	215

1AJK	A/B	214
1AJO	A/B	214
1E4Y	A/B	214
1GTV	A/B	214
1GNW	A/B	211
1EU3	A/B	210
1HW5	A/B	210
1G0S	A/B	209
1GM7	A/B	209
1DJL	A/B	207
1DOW	A/B	205
1H6P	A/B	203
1F6B	A/B	198
1FJR	A/B	195
1AOE	A/B	192
1FBT	A/B	190
1ATZ	A/B	189
1BPL	A/B	189
1CR5	A/B	189
1EX2	A/B	189
1HRU	A/B	188
1D2O	A/B	187
1G2Q	A/B	187
1GXJ	A/B	186
1H1O	A/B	183
1HGX	A/B	183
1F5M	A/B	180
1F3V	A/B	179
1GHE	A/B	177
1AG9	A/B	175

1AOC	A/B	175
1GWY	A/B	175
1ALV	A/B	173
1DVK	A/B	173
1E6C	A/B	173
1BTK	A/B	169
1BO4	A/B	168
1D1G	A/B	168
1AU1	A/B	166
1EPA	A/B	164
1BEB	A/B	162
1EVX	A/B	162
1EXT	A/B	162
1F35	A/B	162
1D1Q	A/B	161
1ALL	A/B	160
1DYO	A/B	160
1DZK	A/B	157
1E7L	A/B	157
1ELK	A/B	157
1EYV	A/B	156
1EM9	A/B	154
1AQZ	A/B	149
1F2T	A/B	149
1F08	A/B	148
1AOH	A/B	147
1EGI	A/B	147
1H97	A/B	147
1GVJ	A/B	146
1EAQ	A/B	140
1F46	A/B	140
1H9S	A/B	140
1DQE	A/B	137
1F7D	A/B	136
1BKZ	A/B	135
1F9Z	A/B	135
1DM9	A/B	133
1FTP	A/B	133
1EMU	A/B	132
1BBH	A/B	131
1ELR	A/B	131
1HPC	A/B	131
1AYO	A/B	130
1COZ	A/B	129
1GY6	A/B	127
1DBW	A/B	126
1EAJ	A/B	126
1ECS	A/B	126
1AKS	A/B	125
1BYF	A/B	125
1DY5	A/B	124
1GU2	A/B	124
1BM9	A/B	122
1D9C	A/B	121
1B2P	A/B	119
1BND	A/B	119
1BHD	A/B	118
1DJ7	A/B	117
1H4X	A/B	117
1H8U	A/B	117
1G8E	A/B	116
1F86	A/B	115
1HXR	A/B	115
1EVH	A/B	112
1F9M	A/B	112
1B0N	A/B	111
1A2P	A/B	110
1AC6	A/B	110
1ECM	A/B	109
1GYO	A/B	109
1CMC	A/B	104
1D4T	A/B	104
1D0Q	A/B	103
1AYA	A/B	101
1CQK	A/B	101
1CQM	A/B	101

Appendix II Biological Protein-protein complexes examined in study with interfaces parameters

12AS_A/12AS_B	1A4Y_A/1A4Y_B	1A7T_A/1A7T_B	1AOR_A/1AOR_B	1AQ6_A/1AQ6_B
1AZT_A/1AZT_B	1B34_A/1B34_B	1B3A_A/1B3A_B	1BQU_A/1BQU_B	1BUH_A/1BUH_B
1CY9_A/1CY9_B	1D09_A/1D09_B	1D0Q_A/1D0Q_B	1DQE_A/1DQE_B	1DQS_A/1DQS_B
1EAJ_A/1EAJ_B	1ECS_A/1ECS_B	1EE8_A/1EE8_B	1ETH_A/1ETH_B	1EUV_A/1EUV_B
1F75_A/1F75_B	1FBT_A/1FBT_B	1FJH_A/1FJH_B	1GPE_A/1GPE_B	1GU7_A/1GU7_B
1H41_A/1H41_B	1H4R_A/1H4R_B	1H54_A/1H54_B	1LFD_A/1LFD_B	1MSP_A/1MSP_B
1QFH_A/1QFH_B	1QOR_A/1QOR_B	1RRP_A/1RRP_B	1XSO_A/1XSO_B	1YCS_A/1YCS_B
1A8U_A/1A8U_B	1AB8_A/1AB8_B	1AC6_A/1AC6_B	1AT3_A/1AT3_B	1AU1_A/1AU1_B
1B5P_A/1B5P_B	1BBH_A/1BBH_B	1BK5_A/1BK5_B	1BYK_A/1BYK_B	1CHM_A/1CHM_B
1D1G_A/1D1G_B	1D2O_A/1D2O_B	1D9C_A/1D9C_B	1DVK_A/1DVK_B	1E19_A/1E19_B
1EEJ_A/1EEJ_B	1EG5_A/1EG5_B	1EG9_A/1EG9_B	1EX2_A/1EX2_B	1EZI_A/1EZI_B
1FP3_A/1FP3_B	1FSS_A/1FSS_B	1FTP_A/1FTP_B	1GV3_A/1GV3_B	1GVE_A/1GVE_B
1H6P_A/1H6P_B	1HJR_A/1HJR_C	1HPC_A/1HPC_B	1ONE_A/1ONE_B	1PDK_A/1PDK_B
1SMP_I/1SMP_A	1SPU_A/1SPU_B	1STF_E/1STF_I	2AE2_A/2AE2_B	2HHM_A/2HHM_B
1AD1_A/1AD1_B	1AK4_A/1AK4_D	1ALV_A/1ALV_B	1AVW_A/1AVW_B	1CQK_A/1CQK_B
1BKD_R/1BKD_S	1BKP_A/1BKP_B	1BKZ_A/1BKZ_B	1CMC_A/1CMC_B	1CQ3_A/1CQ3_B
1DDZ_A/1DDZ_B	1DKU_A/1DKU_B	1DLE_A/1DLE_B	1E5X_A/1E5X_B	4SGB_I/4SGB_E
1EGI_A/1EGI_B	1EI1_A/1EI1_B	1EKE_A/1EKE_B	1F34_A/1F34_B	1F6Y_A/1F6Y_B
1G0H_A/1G0H_B	1G60_A/1G60_B	1G8E_A/1G8E_B	1GYO_A/1GYO_B	1H3F_A/1H3F_B
1I2M_A/1I2M_B	1ISA_A/1ISA_B	1ITB_A/1ITB_B	1QAE_A/1QAE_B	1QAV_A/1QAV_B
1TAB_I/1TAB_E	1TGS_I/1TGS_Z	1TRK_A/1TRK_B	2PFL_A/2PFL_B	2PTC_I/2PTC_E
1AOC_A/1AOC_B	1AOH_A/1AOH_B	1AOM_A/1AOM_B	1ASO_A/1ASO_B	1AUO_A/1AUO_B
1BO1_A/1BO1_B	1BO4_A/1BO4_B	1BPL_A/1BPL_B	1BVN_T/1BVN_P	1CI9_A/1CI9_B
1DMH_A/1DMH_B	1DN1_A/1DN1_B	1DOR_A/1DOR_B	1DQZ_A/1DQZ_B	1E5R_A/1E5R_B
1EMV_A/1EMV_B	1EPA_A/1EPA_B	1EQ9_A/1EQ9_B	1EX0_A/1EX0_B	1F0K_A/1F0K_B
1GDE_A/1GDE_B	1GNW_A/1GNW_B	1GOI_A/1GOI_B	1GUX_A/1GUX_B	1GXJ_A/1GXJ_B
1JTD_A/1JTD_B	1KAC_A/1KAC_B	1KPE_A/1KPE_B	1NSE_A/1NSE_B	1PP2_L/1PP2_R
1VLT_A/1VLT_B	1VOK_A/1VOK_B	1WQ1_R/1WQ1_G	1ZBD_A/1ZBD_B	2PCB_A/2PCB_B

Appendix III. Non-biological Protein-protein complexes examined in study with interfaces parameters

1A04_A/1A04_B	1DK7_A/1DK7_B	1FJR_A/1FJR_B	1GWY_A/1GWY_B	1IBQ_A/1IBQ_B
1AGJ_A/1AGJ_B	1DKL_A/1DKL_B	1FMJ_A/1FMJ_B	1GXM_A/1GXM_B	1ICP_A/1ICP_B
1AJK_A/1AJK_B	1DNP_A/1DNP_B	1FMT_A/1FMT_B	1GXY_A/1GXY_B	1IK7_A/1IK7_B
1AMU_A/1AMU_B	1DVG_A/1DVG_B	1FNN_A/1FNN_B	1H03_P/1H03_Q	1IM8_A/1IM8_B
1AOE_A/1AOE_B	1DY5_A/1DY5_B	1FSL_A/1FSL_B	1H0B_A/1H0B_B	1IN0_A/1IN0_B
1AQZ_A/1AQZ_B	1DZK_A/1DZK_B	1FVR_A/1FVR_B	1H1O_A/1H1O_B	1IO7_A/1IO7_B
1ATL_A/1ATL_B	1E0X_A/1E0X_B	1FZY_A/1FZY_B	1H3G_A/1H3G_B	1IOO_A/1IOO_B
1B3U_A/1B3U_B	1E30_A/1E30_B	1G1B_A/1G1B_B	1H4P_A/1H4P_B	1IQ4_A/1IQ4_B
1BF6_A/1BF6_B	1E6C_A/1E6C_B	1G1K_A/1G1K_B	1H6G_A/1H6G_B	1IT2_A/1IT2_B
1BGE_A/1BGE_B	1E6F_A/1E6F_B	1G4M_A/1G4M_B	1H7S_A/1H7S_B	1IU1_A/1IU1_B
1BIR_A/1BIR_B	1E8C_A/1E8C_B	1G61_A/1G61_B	1H8U_A/1H8U_B	1IWM_A/1IWM_B
1C0E_A/1C0E_B	1E9N_A/1E9N_B	1GEQ_A/1GEQ_B	1HA3_A/1HA3_B	1IYK_A/1IYK_B
1CQM_A/1CQM_B	1EAQ_A/1EAQ_B	1GG4_A/1GG4_B	1HJZ_A/1HJZ_B	1IZ5_A/1IZ5_B
1CQX_A/1CQX_B	1ELK_A/1ELK_B	1GHE_A/1GHE_B	1HM6_A/1HM6_B	1J2F_A/1J2F_B
1CZF_A/1CZF_B	1ETP_A/1ETP_B	1GIQ_A/1GIQ_B	1HPL_A/1HPL_B	1J6R_A/1J6R_B
1D1Q_A/1D1Q_B	1EU3_A/1EU3_B	1GOU_A/1GOU_B	1HX3_A/1HX3_B	1J7J_A/1J7J_B
1D7J_A/1D7J_B	1F0X_A/1F0X_B	1GQP_A/1GQP_B	1HXR_A/1HXR_B	1J83_A/1J83_B
1DBW_A/1DBW_B	1F2K_A/1F2K_B	1GT6_A/1GT6_B	1HY5_A/1HY5_B	1J96_A/1J96_B
1DBX_A/1DBX_B	1F35_A/1F35_B	1GUD_A/1GUD_B	1I19_A/1I19_B	1J97_A/1J97_B
1DJX_A/1DJX_B	1F9M_A/1F9M_B	1GV4_A/1GV4_B	1I7K_A/1I7K_B	1JBB_A/1JBB_B
1JCL_A/1JCL_B	1JH6_A/1JH6_B	1JPA_A/1JPA_B	1JSS_A/1JSS_B	
1JFR_A/1JFR_B	1JIH_A/1JIH_B	1JQE_A/1JQE_B	1JU2_A/1JU2_B	
1JFU_A/1JFU_B	1JJT_A/1JJT_B	1JR2_A/1JR2_B	1JVA_A/1JVA_B	